Details of first Dataset (application_data.csv)

1. Shape – **(307511,122)**
2. Column count with more than 40% null values – **49**
3. Column count with null percentage <40% - **18**

Deleted all columns with null percentage more than 40% as it is a significantly huge number.

Analysing columns with null percentage >0<40

1. OCCUPATION_TYPE – 31%

These null values mostly correspond to Pensioners and since the null value count is high, we can create a separate category for them “Unknown”.

2. EXT_SOURCE_3 – 19.8%

Mean and median values for this column are very close. We can use median value here for imputation as there's a big diff between min and max values. So using a median would take care of them but since the null percentage is around 20% so we can leave these values as is.

3. AMT_REQ_CREDIT_BUREAU_YEAR – 13.5%

Since there are a lot entries with count as 0 therefore we cannot imply that Nan here stands for zero calls/enquiries made. By looking at the value counts records, there is not a very big difference between value counts for '0.0' and '1.0' and '2.0'so here to be safe we can pick median but since the number of null values is high we can leave them as it is.

4. AMT_REQ_CREDIT_BUREAU_MON – 13.5%

The missing values here can be imputed with the mode value i.e 0.0 as the frequency of this value is too high.

Also till 75% the value is 0.0000.

5. NAME_TYPE_SUITE – 0.4%

clearly unaccompanied has been the most frequent choice of customers, so we can impute the mising values with this category type.

For columns like **AMT_GOODS_PRICE** and **AMT_ANNUITY**, the null percentages are very low 0.090403 and 0.003902 respectively and hence these records can be dropped.

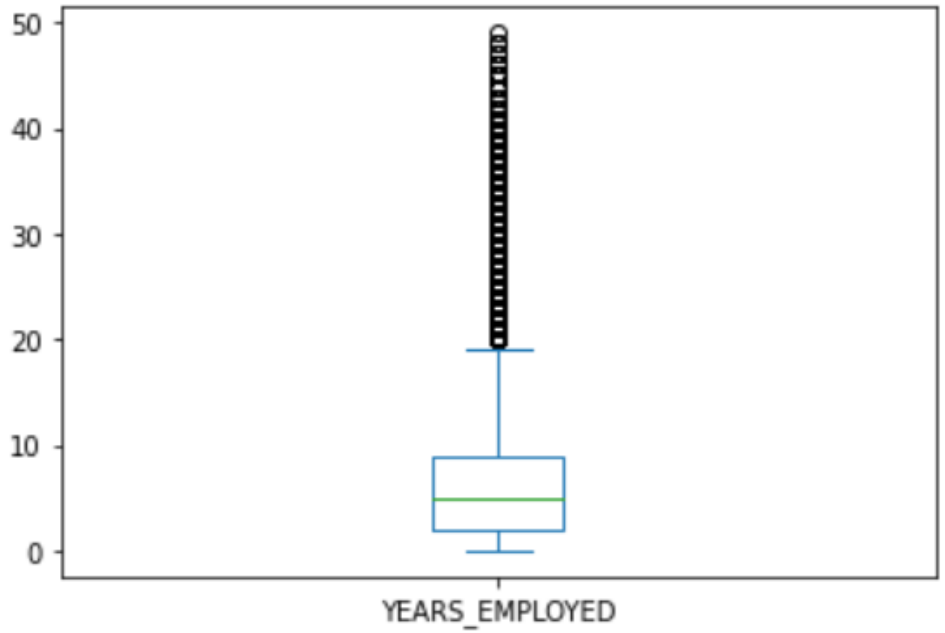
OUTLIER (UNIVARIATE ANALYSIS)

1. YEARS_EMPLOYED COLUMN

Inference:

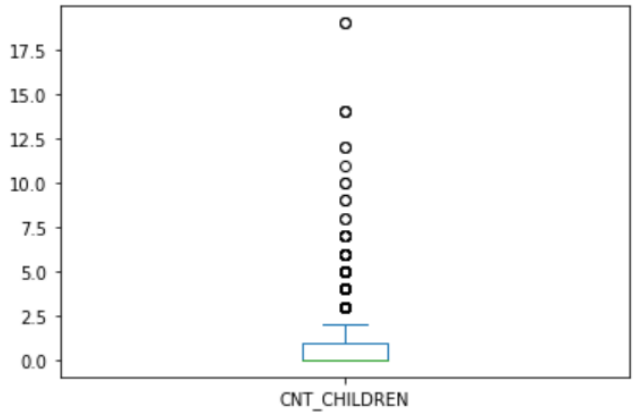
There is clearly an outlier value for years of employment : 1001 but for all these records the employment type is Pensioner.

This age group holds a very big count value and therefore we should cap this at some reasonable value



OUTLIER (UNIVARIATE ANALYSIS)

- 2. CNT_CHILDREN COLUMN



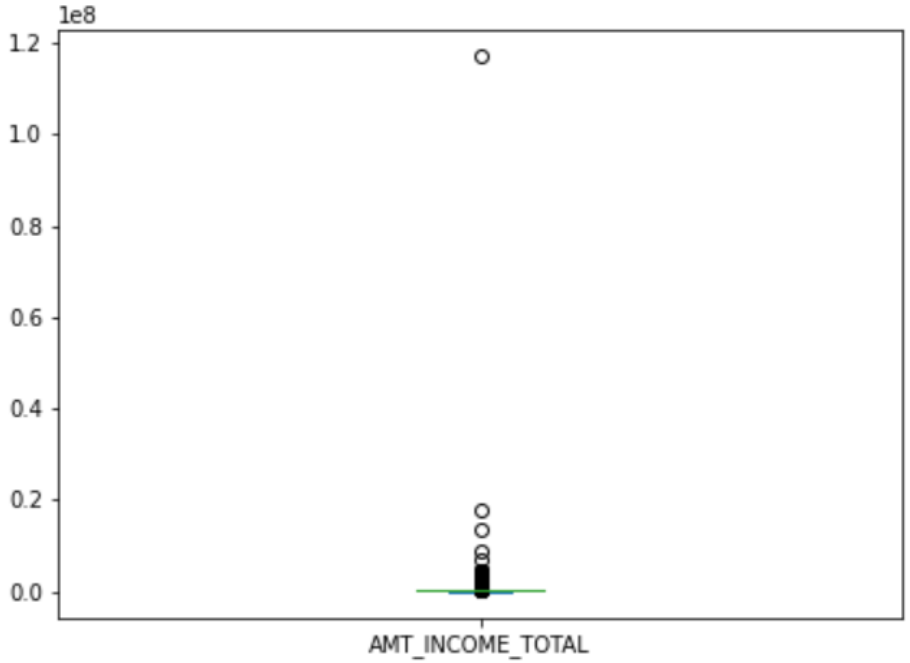
	CNT_CHILDREN	AGE_IN_YEARS	NAME_FAMILY_STATUS
12615	8	42	Married
23881	9	30	Single / not married
34545	11	47	Married
80948	12	39	Married
132585	10	31	Married
154317	8	31	Married
155369	19	30	Single / not married
171125	12	38	Married
176011	14	49	Separated
183878	14	56	Married
186820	10	41	Married
265784	19	28	Single / not married
267998	14	42	Married
276768	9	40	Civil marriage

Inference:

Applicants with more than 7 children are very low in number and they are clearly outliers. Their ages are in mostly in the age group 30-40. Some are even single and have 19 children which is a very rare occurence. We can drop these records.

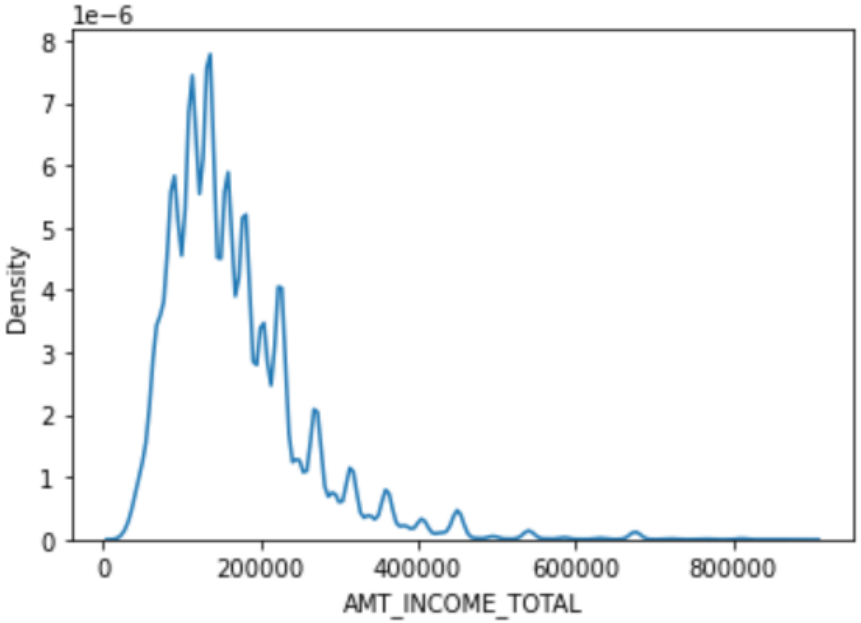
OUTLIER (UNIVARIATE ANALYSIS)

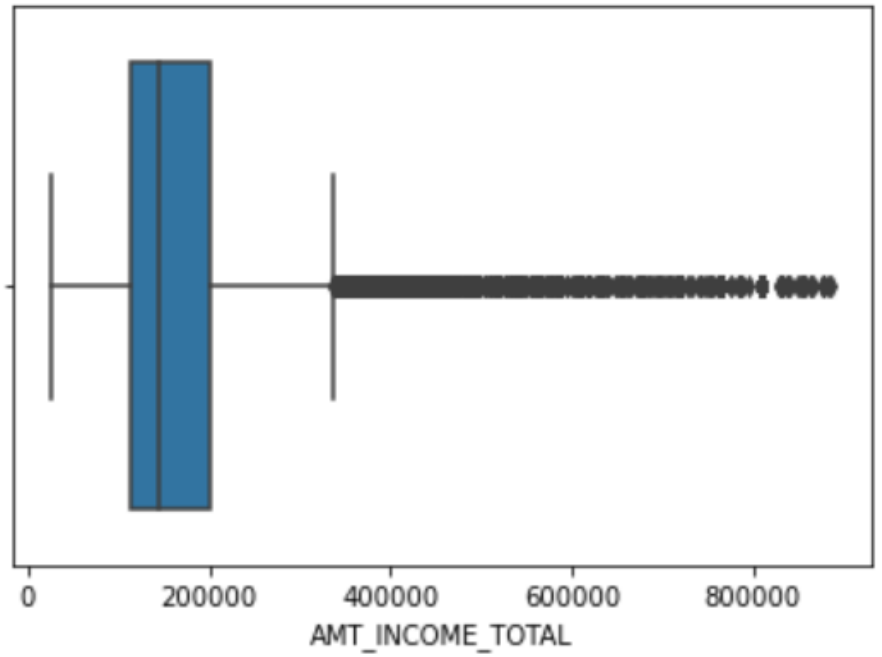
- 3. AMT_INCOME_TOTAL COLUMN



There's a difference between 99th and 99.9th percentile. There's a remarkable diff between the 99.9th and the max value and this value is clearly an outlier. Let's explore this even further.

For values < 900000 :



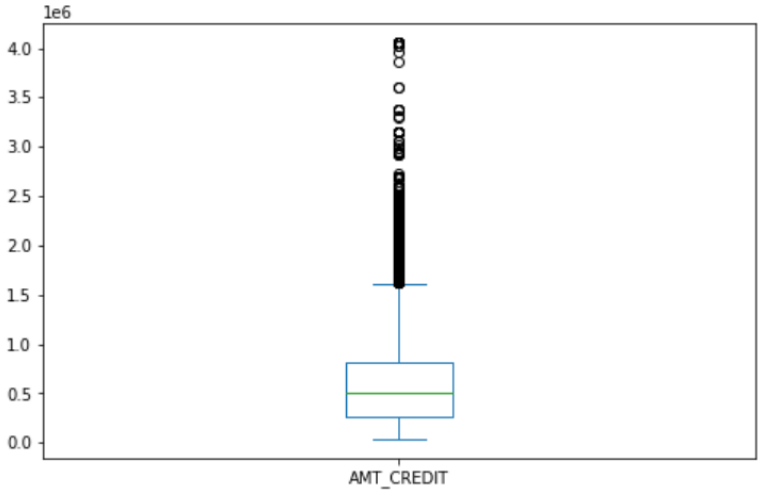


Inference:

after the max value in the above box plot there's a huge chunk of values outside the upper fence but they are continuous values and not very far away from the upper fence. These can be though of as people with higher incomes but clearly applicants with income >900000 are outliers. We should drop these values as these are too high and will disturb our analysis or not can be left as it is but not considered during analysis.

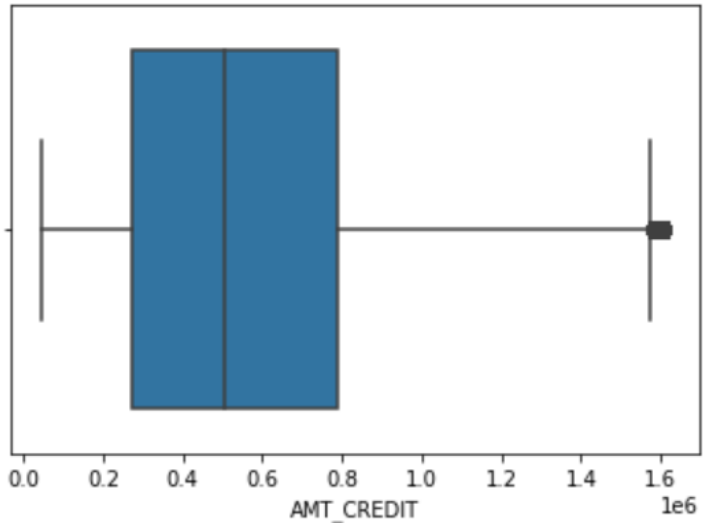
OUTLIER (UNIVARIATE ANALYSIS)

4. AMT_CREDIT COLUMN



For values less than the value of upper fence

Q3 + 1.5 * IQR : 1616625.0

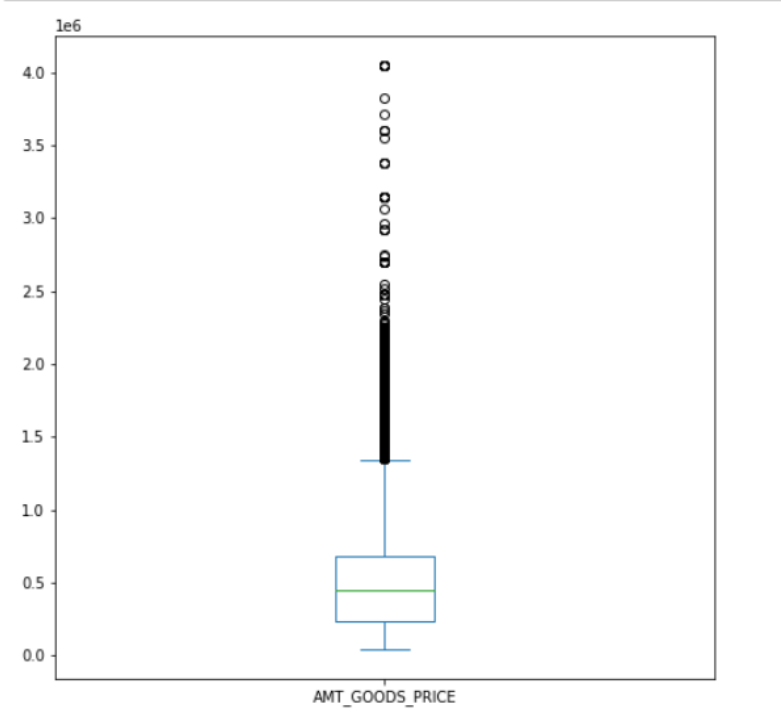


Inference :

Clearly before this max value there are no outliers and all the values beyond max i.e beyond 95th percentile are outliers and should not be considered for analysis.

OUTLIER (UNIVARIATE ANALYSIS)

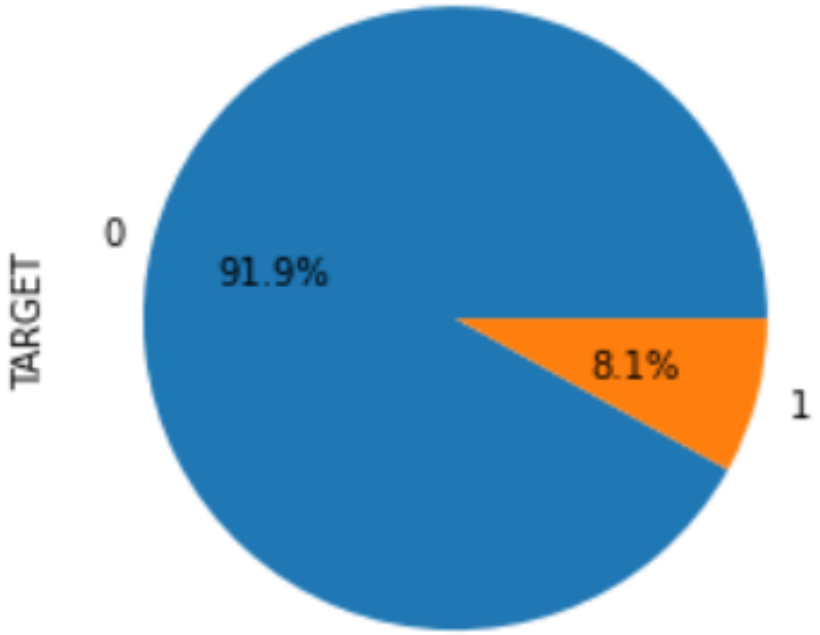
5. AMT_GOODS_PRICE COLUMN



Inference :

Beyond the upper fence, there is a chunk but attached the fence. Beyond the 99th percentile we have outliers and they must be ignored or capped.

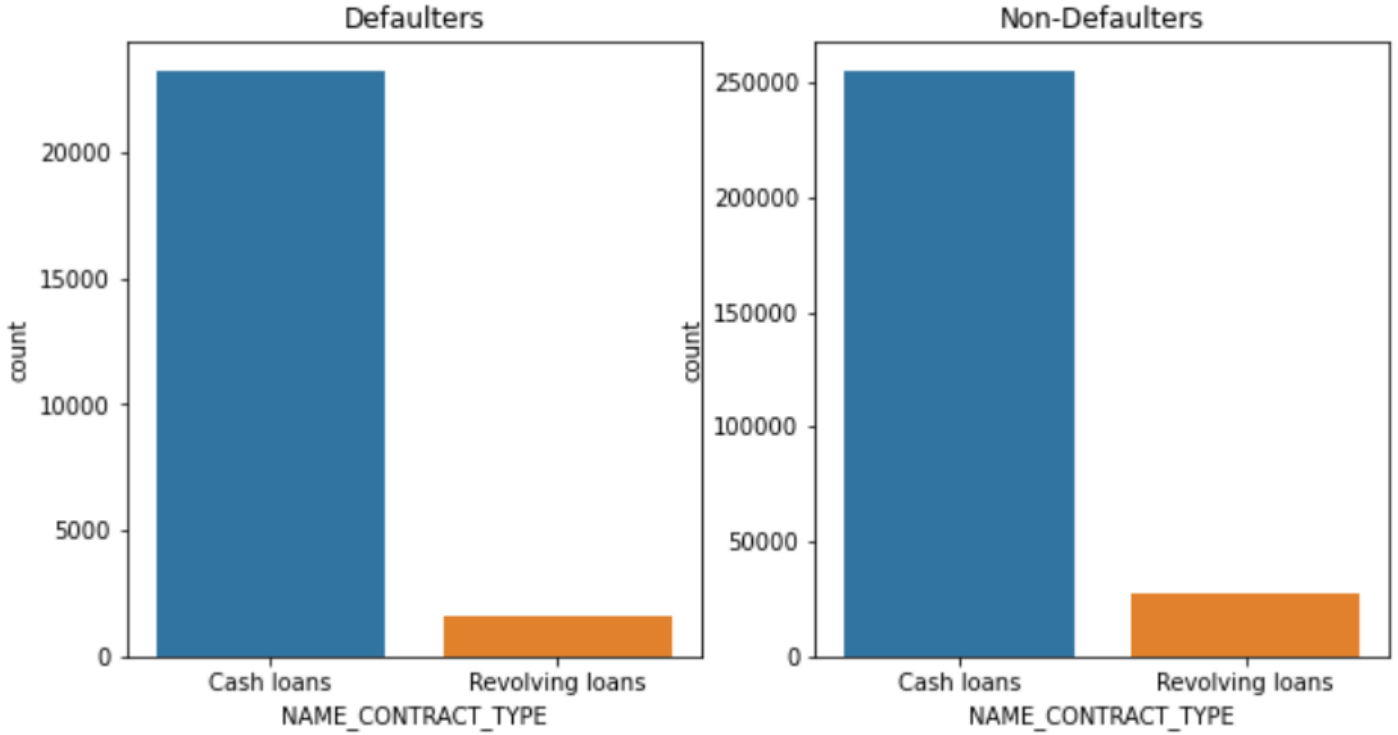
IMBALANCE PERCENTAGE



Around 8% applicants are defaulters and 92% non- defaulters

ANALYSING THE TWO SUBSETS DIVIDED BASED ON TARGET VARIABLE

- Contract Type



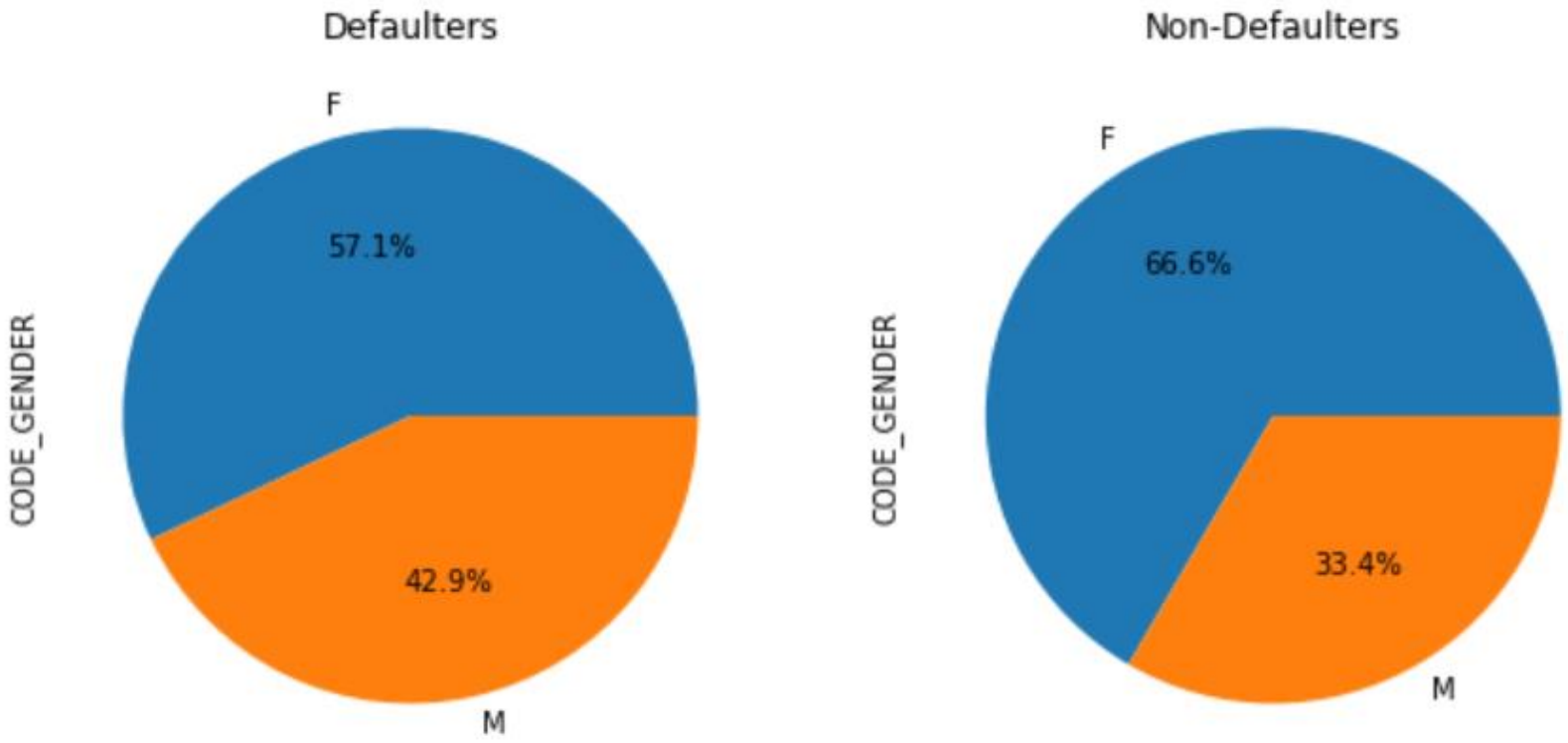
Inference:

Not much difference between defaulters and non defaulters based on different contract types. There's a very slight difference where Revolving Loan type has slightly less defaulters.

Gender Variable

Inference:

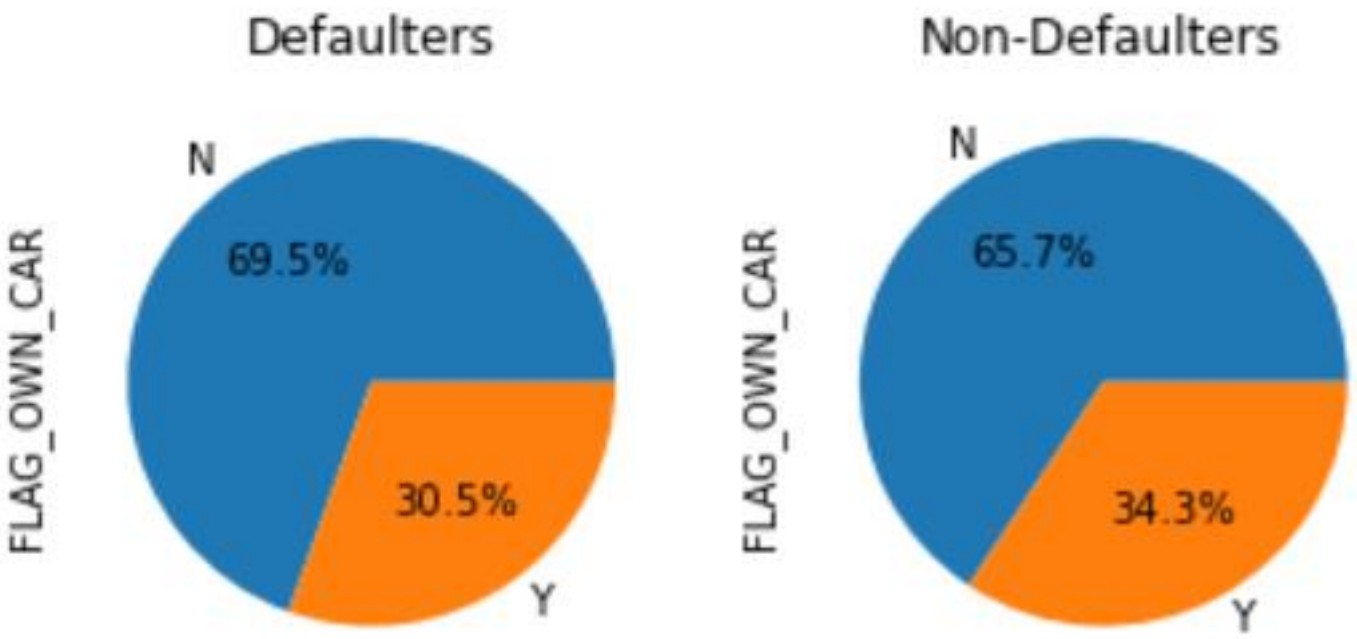
Not much difference but they are about 10% more Males in the defaulters category.



Owning a car

Inference:

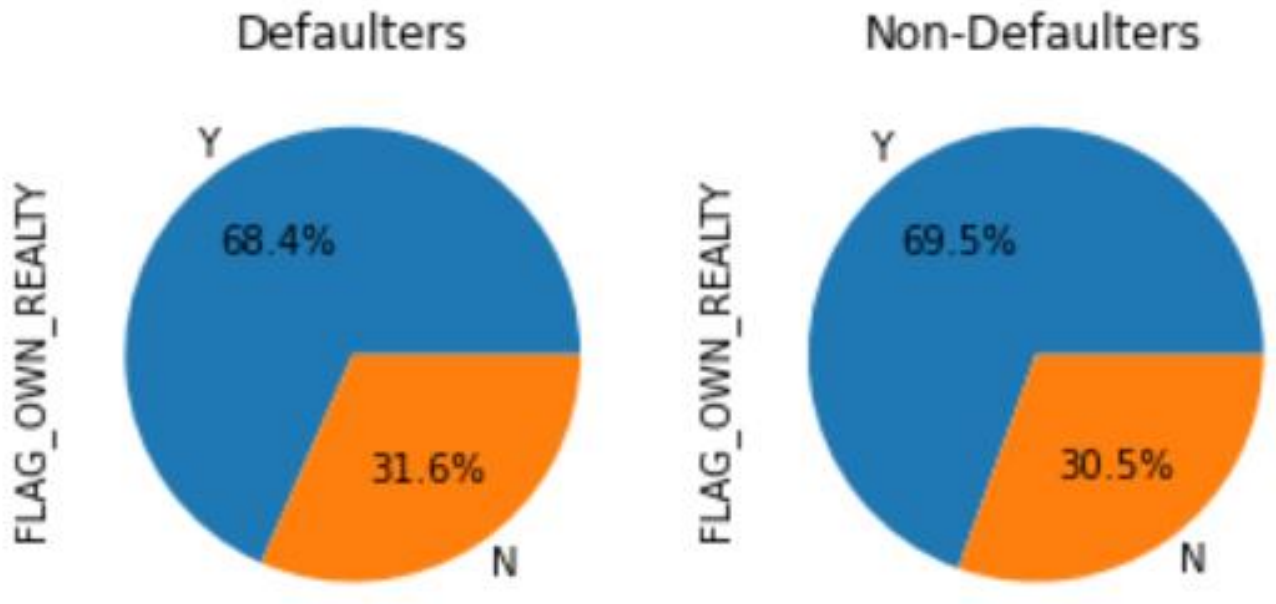
This does not provide any strong evidence of people not being able to pay on time if they own a car.



Owning a Realty

Inference:

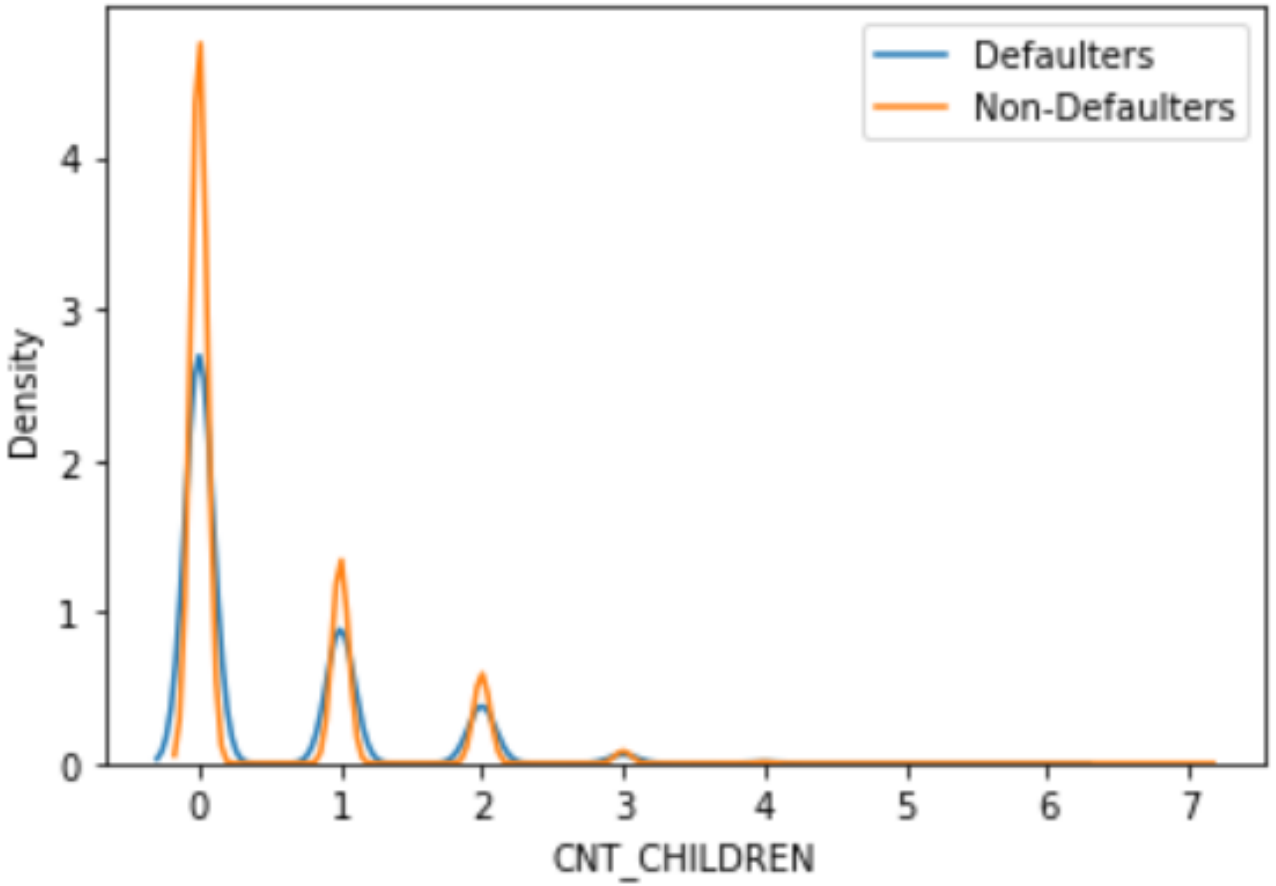
Owning a Realty does not necessarily imply that applicants would not be able to make payments on time.



Number of children

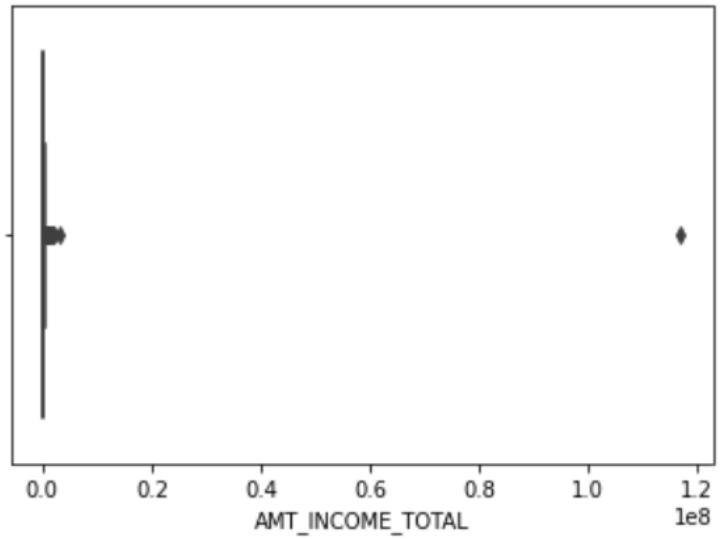
Inference:

There a lot more number of clients with 0 children who make payments on time.

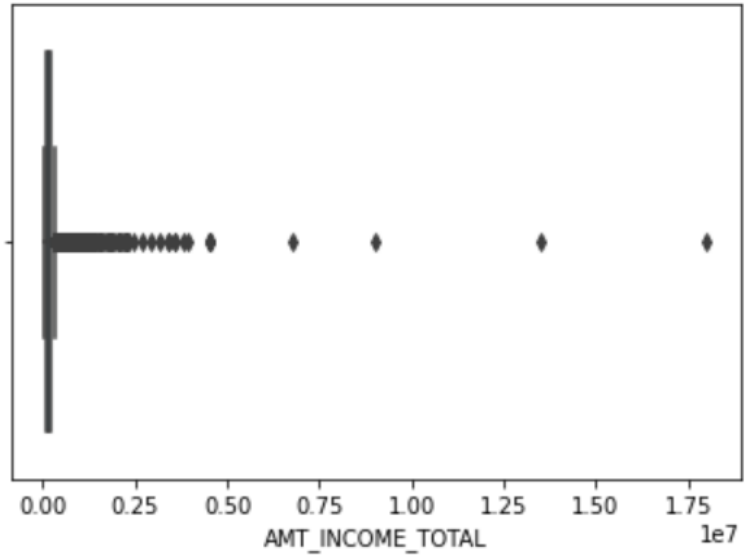


Total Income

Defaulters:



Non-Defaulters:



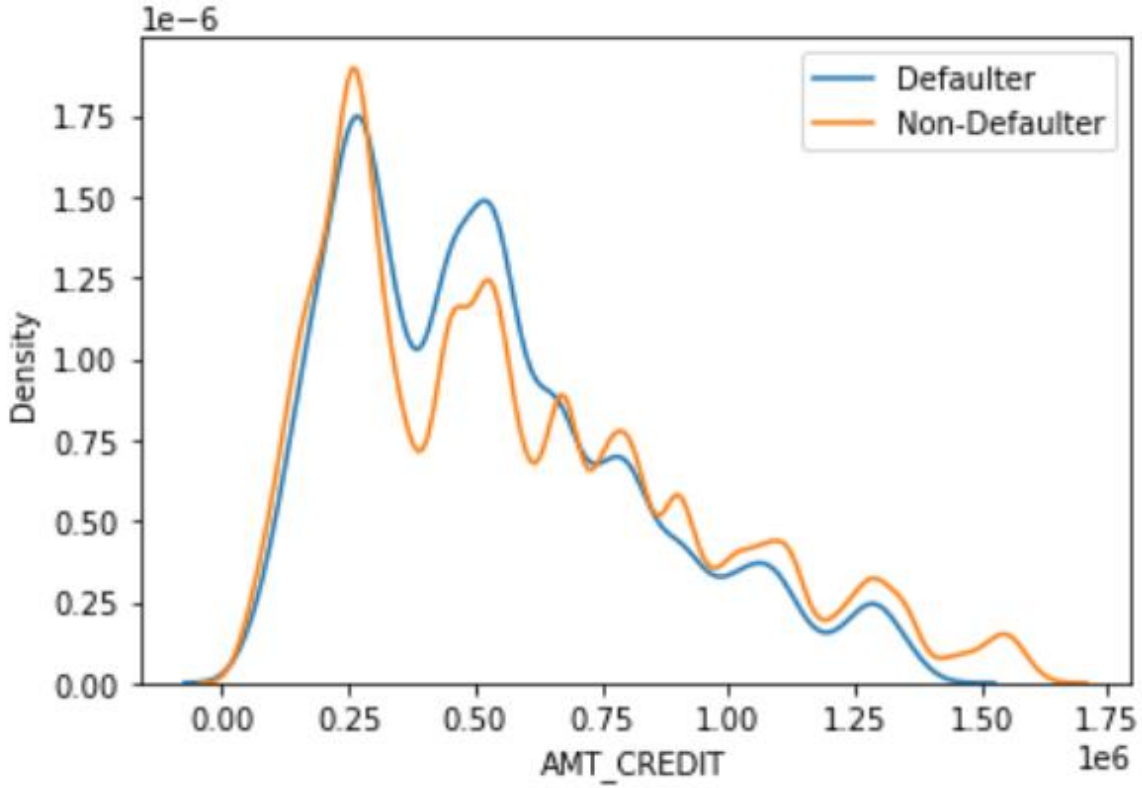
Inference:

Looking at the distplot, we can say that clients having slightly higher incomes tend to pay installments on time.

Credit Amount

Inference:

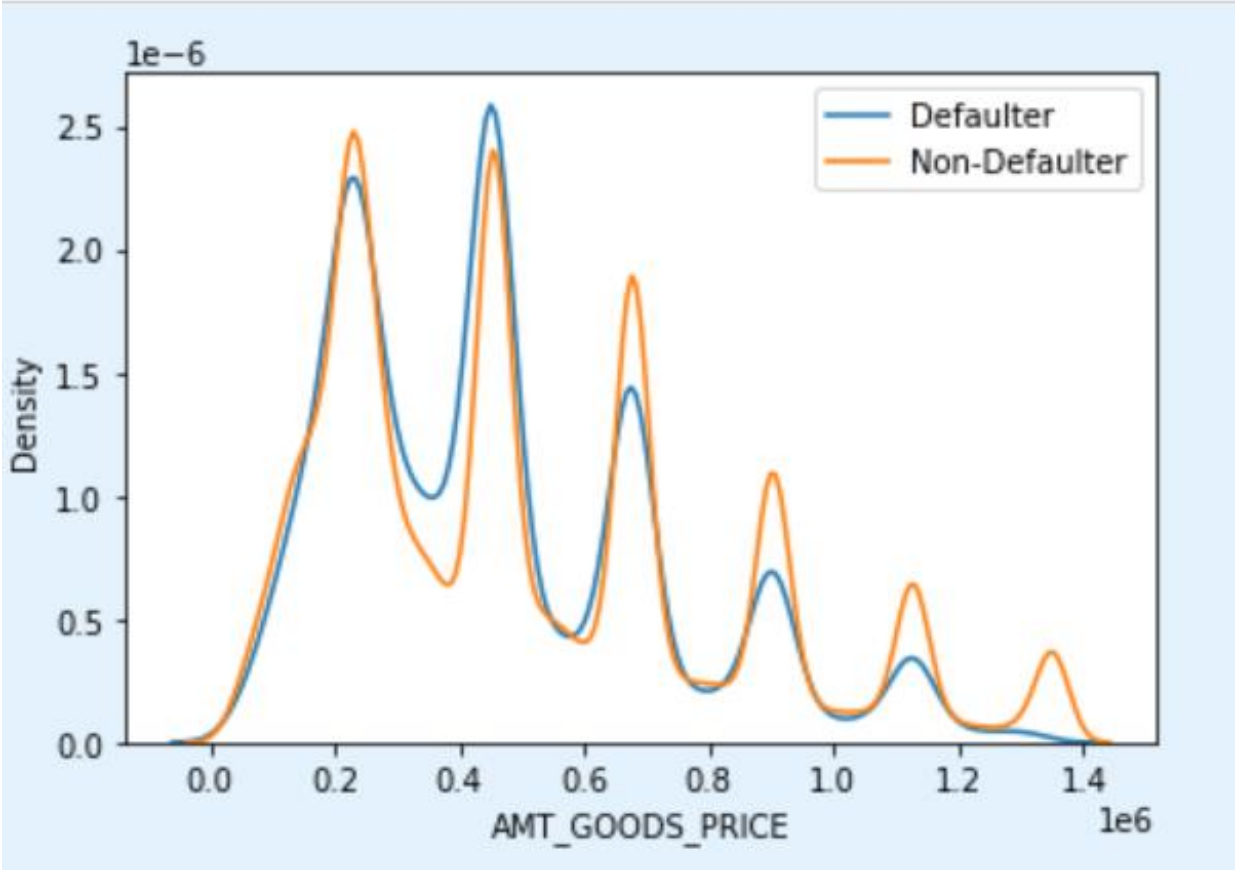
For credited amounts between 400,000 & 650,000 approx there are more clients who have defaulted on the payment. Also, although there are a lot spikes but beyond 750,000 there are more clients who pay on time or we can say that clients who take larger loans mostly pay on time.



Goods Price

Inference:

For goods with values higher than approx 600,000 there are less defaulters.

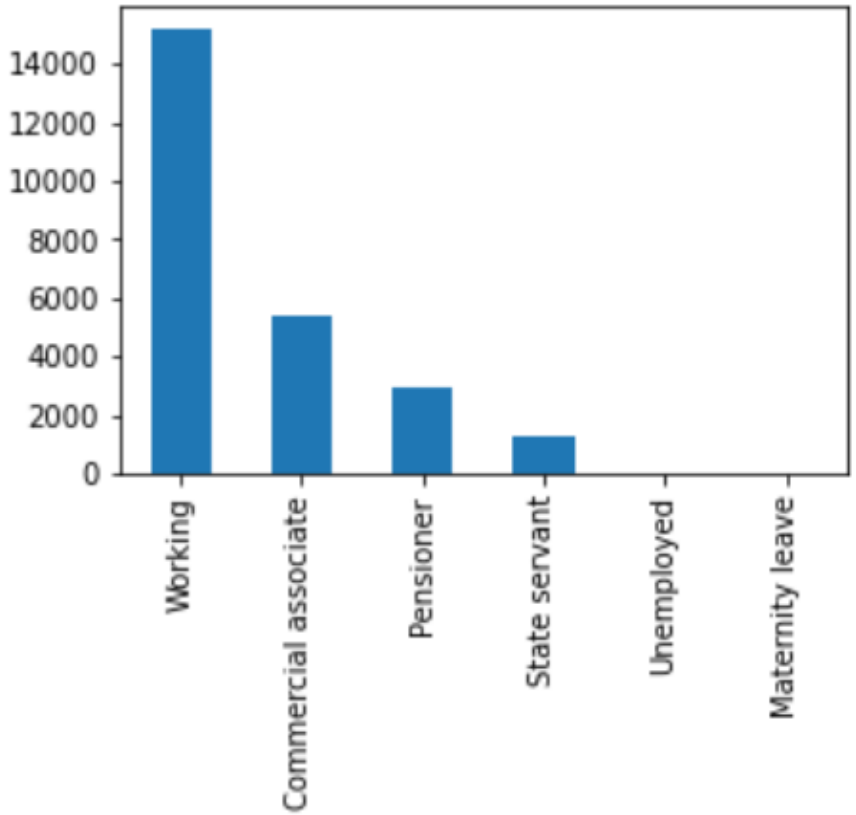


Income type

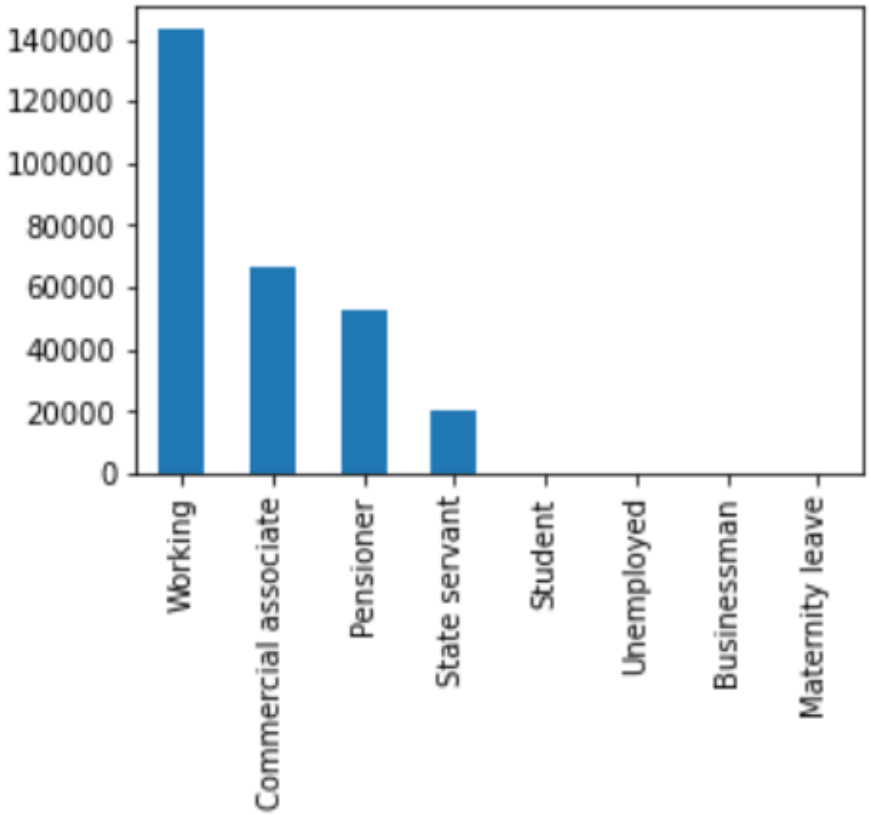
Inference:

Commercial Associates and Pensioners may on time.

Defaulters:



Non-Defaulters:

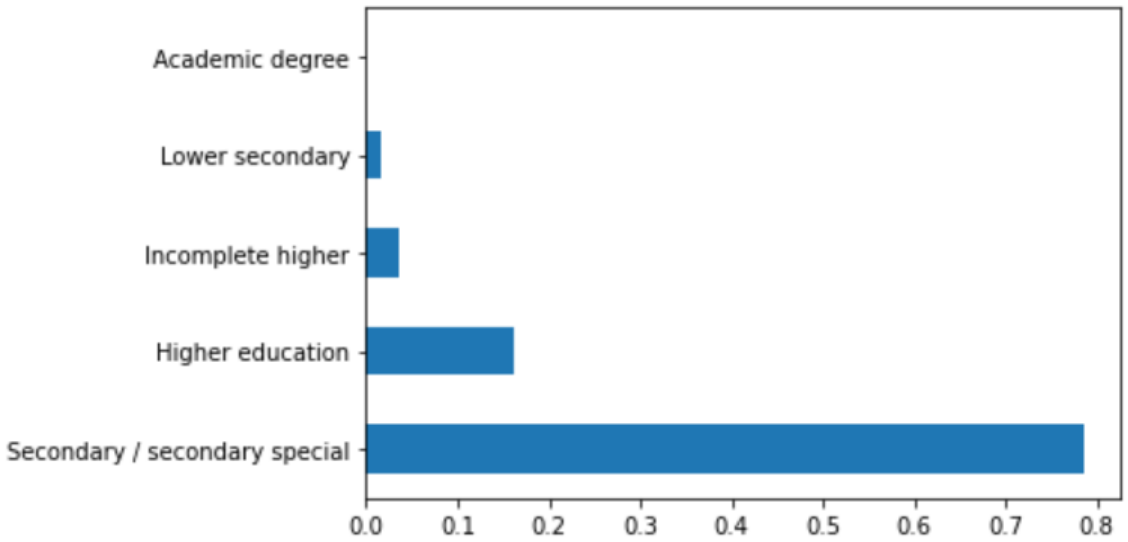


Education Type

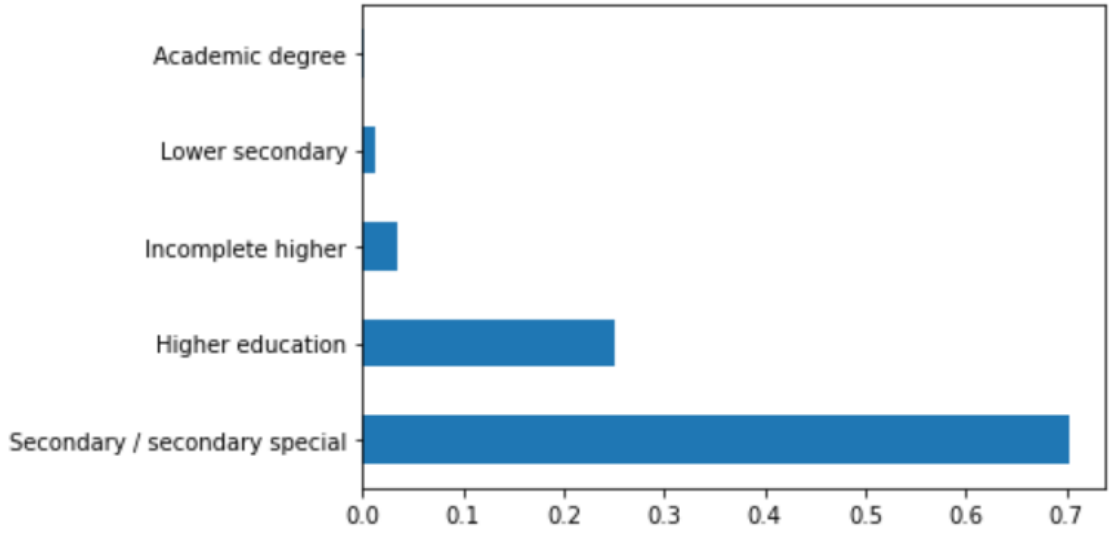
Inference:

Higher Education category make on time payments.

Defaulters:



Non Defaulters:

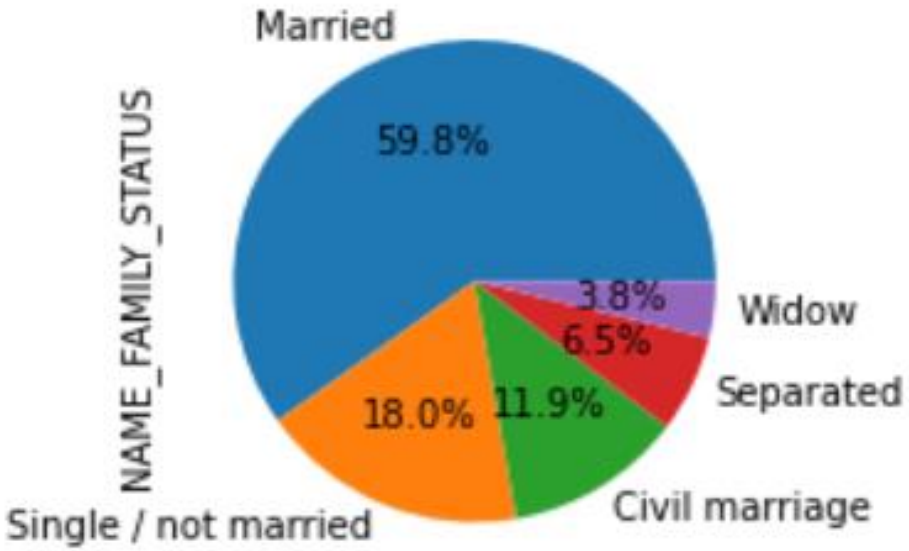


Marital/Family Status

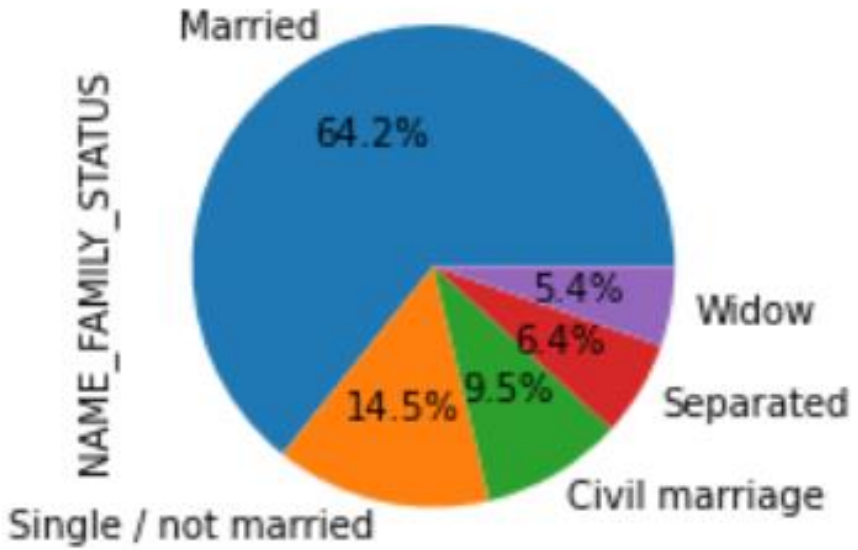
Inference:

This is not a strong observation but overall married clients and widows do on time payment and single/not married clients face difficulties in making on time payments.

Defaulters:



Non-Defaulters:

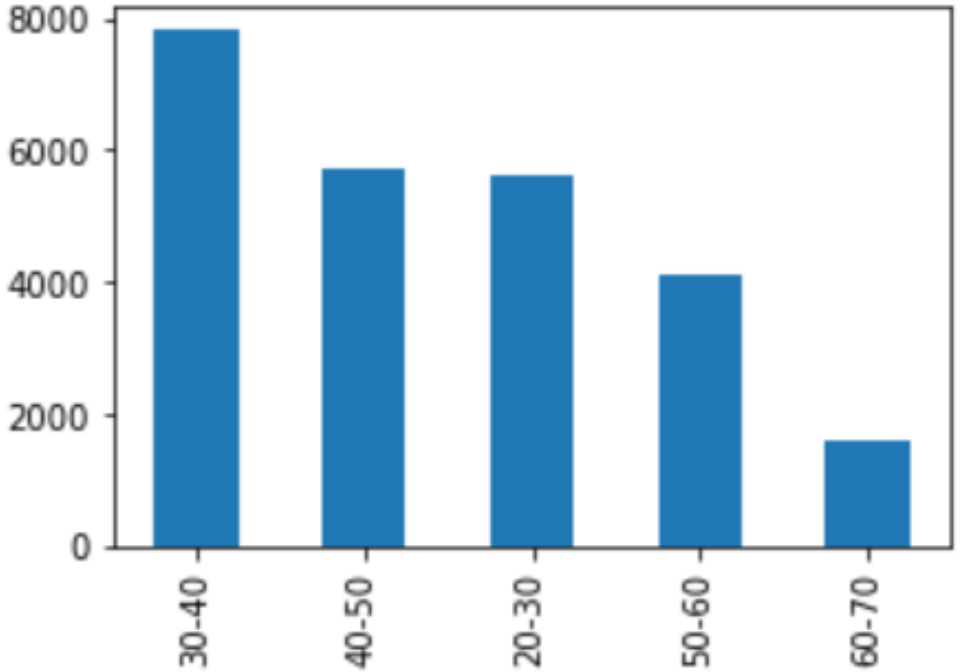


Age-Group

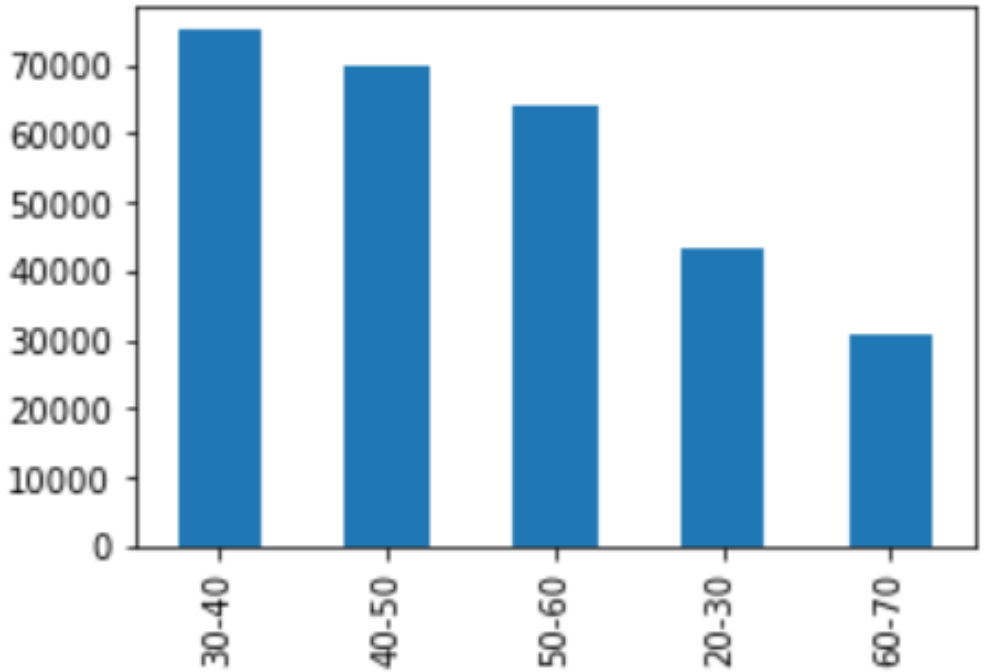
Inference:

Age group 40 and above makes payments on time. 20-30 age group has more defaulters.

Defaulters:



Non-Defaulters:

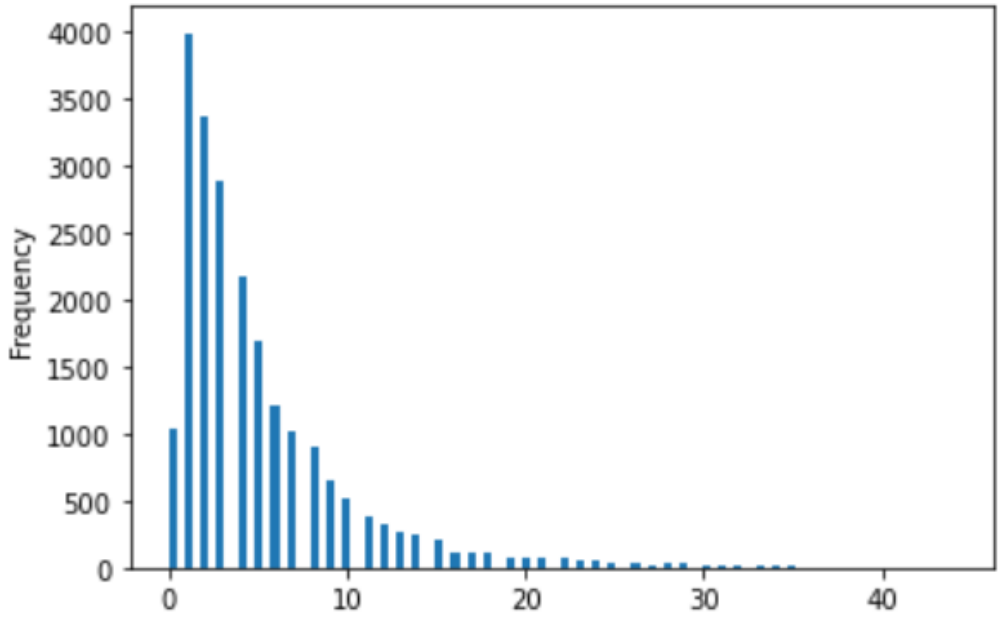


Employment Years

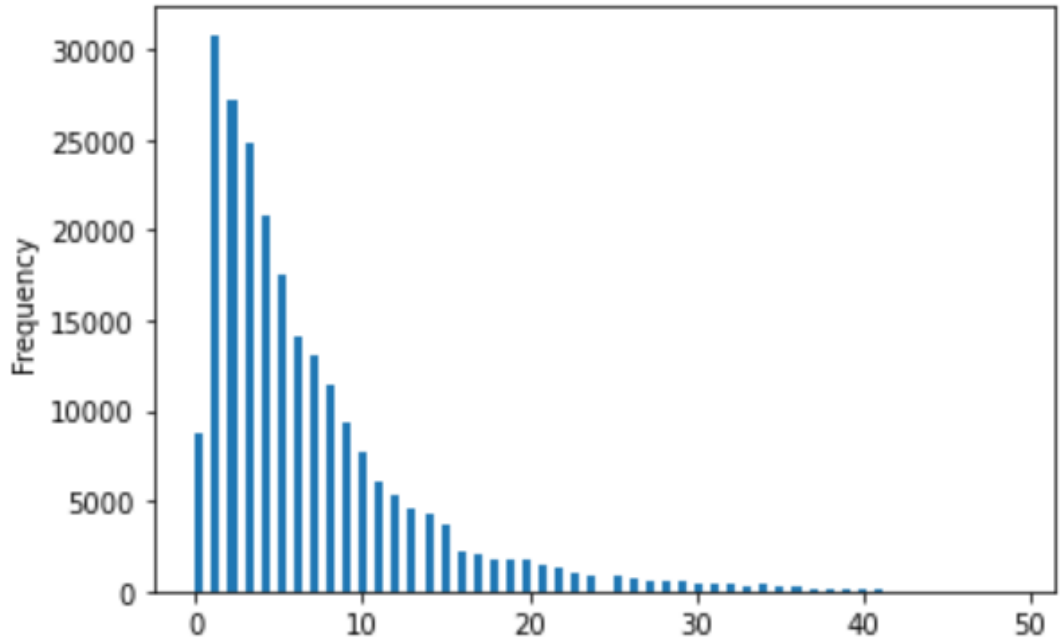
Inference:

Clients who have more years of work experience make on time payments.

Defaulters:



Non-Defaulters:

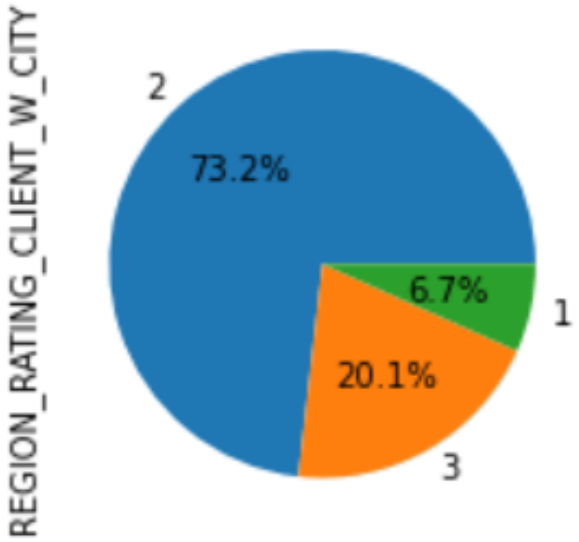


City Wise Region Rating

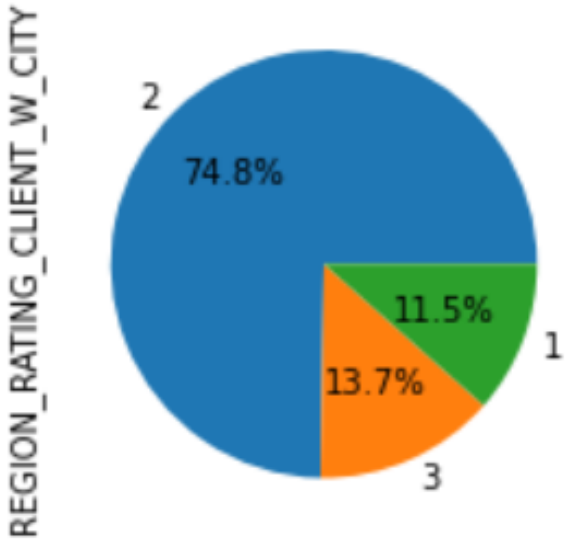
Inference:

Clients from Region rated 3 have more defaulters and region 1 seems to do better.

Defaulters:



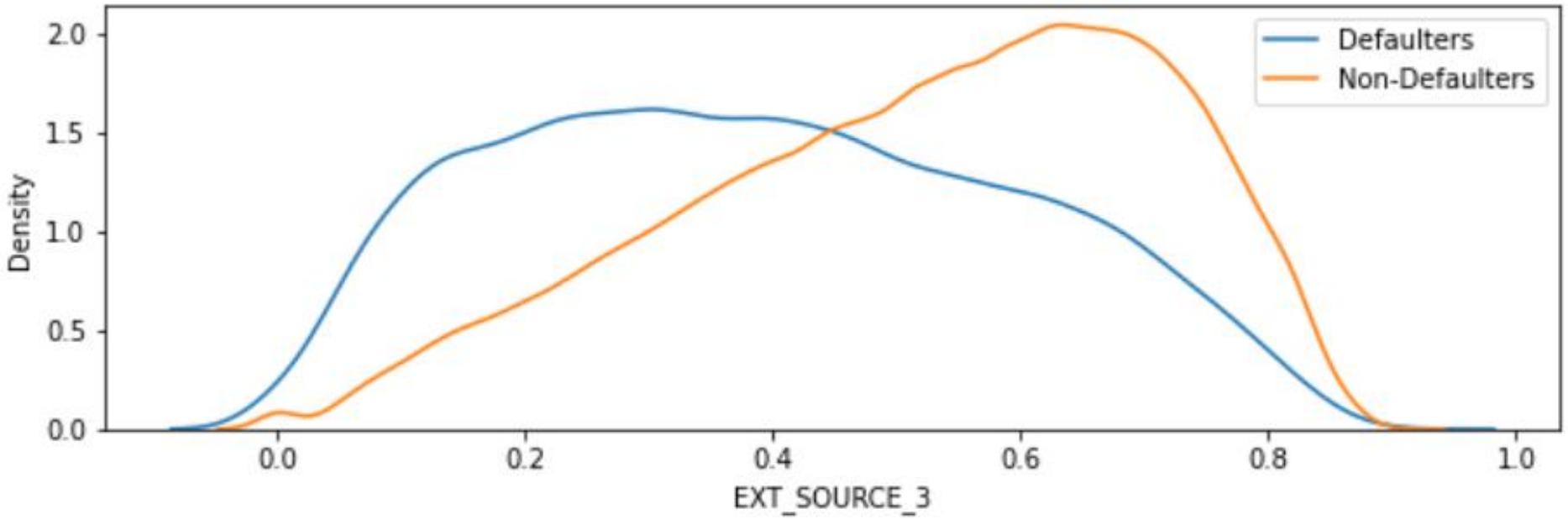
Non-Defaulters:



Score from External source-3

Inference:

Defaulters have low score.



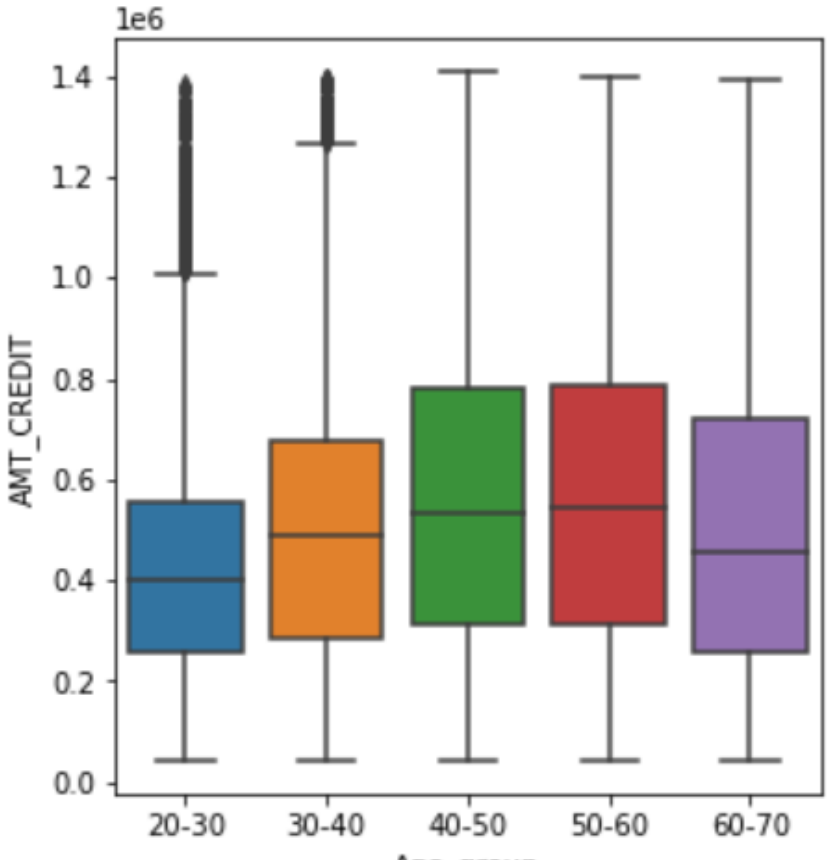
BIVARIATE ANALYSIS

Age_Group with Credit Amount

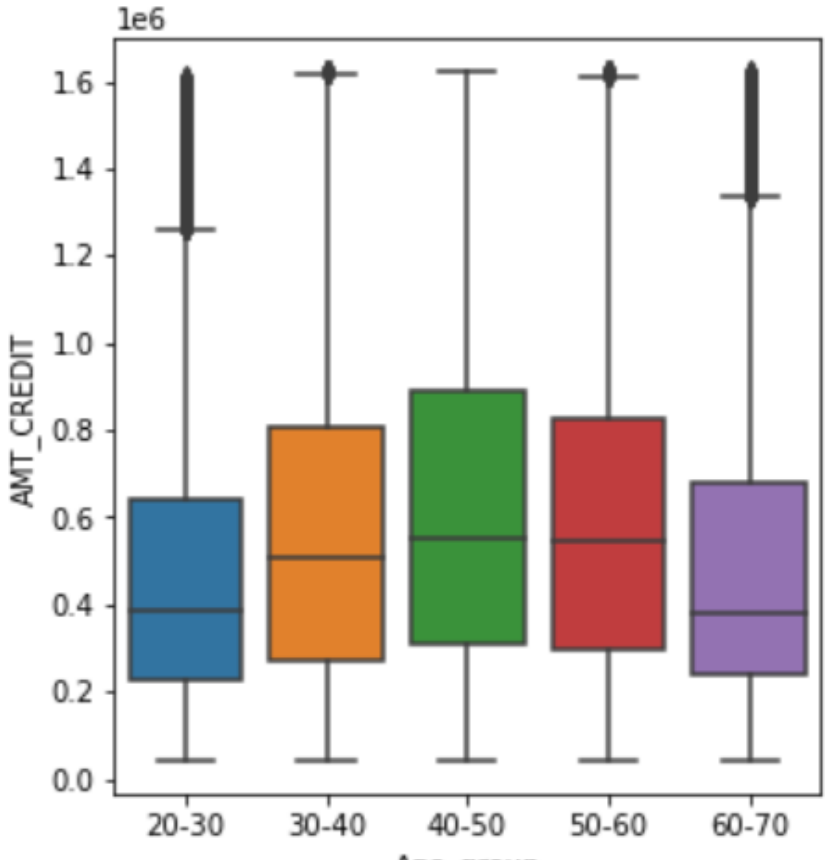
Inference:

Clients in the age group of 40-50 have higher loan credit amounts and make payment on time**.**

Defaulters:



Non-Defaulters:

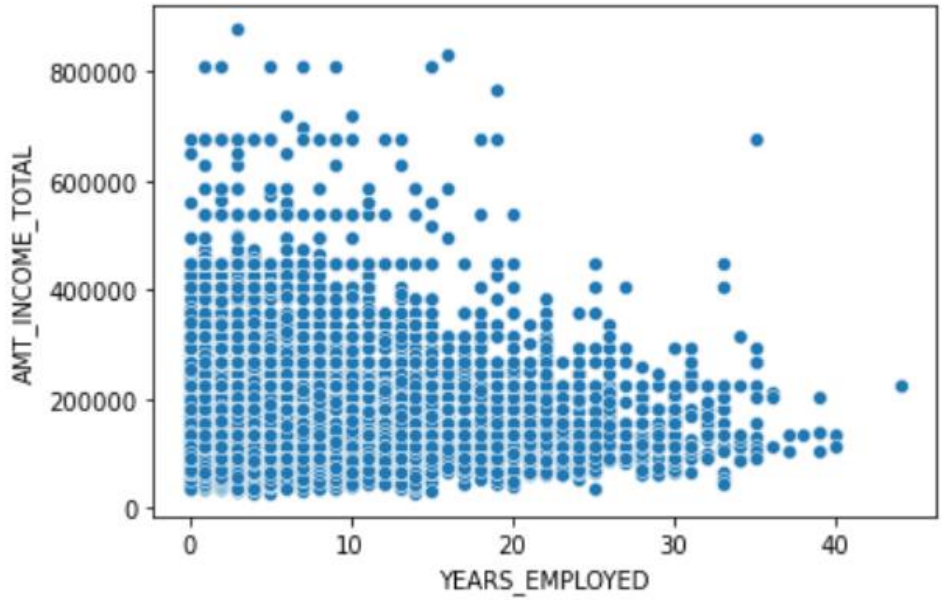


YEARS_EMPLOYED + AMT_INCOME_TOTAL

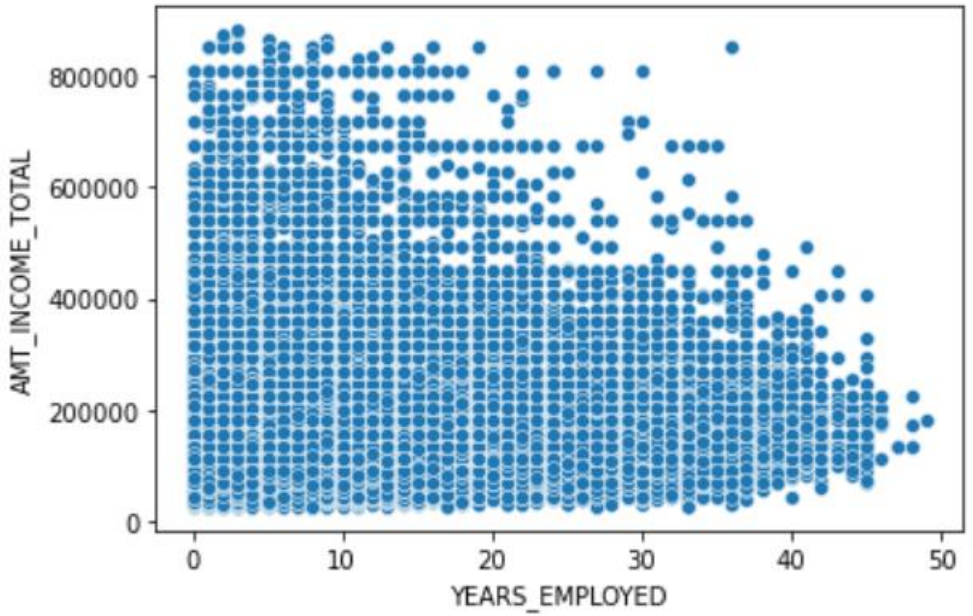
Inference:

clients who have more work experience have higher incomes and therefore are able to make payments in time whereas clients with less work ex have low incomes and therefore fail to make payment on time.

Defaulters:



Non-Defaulters:

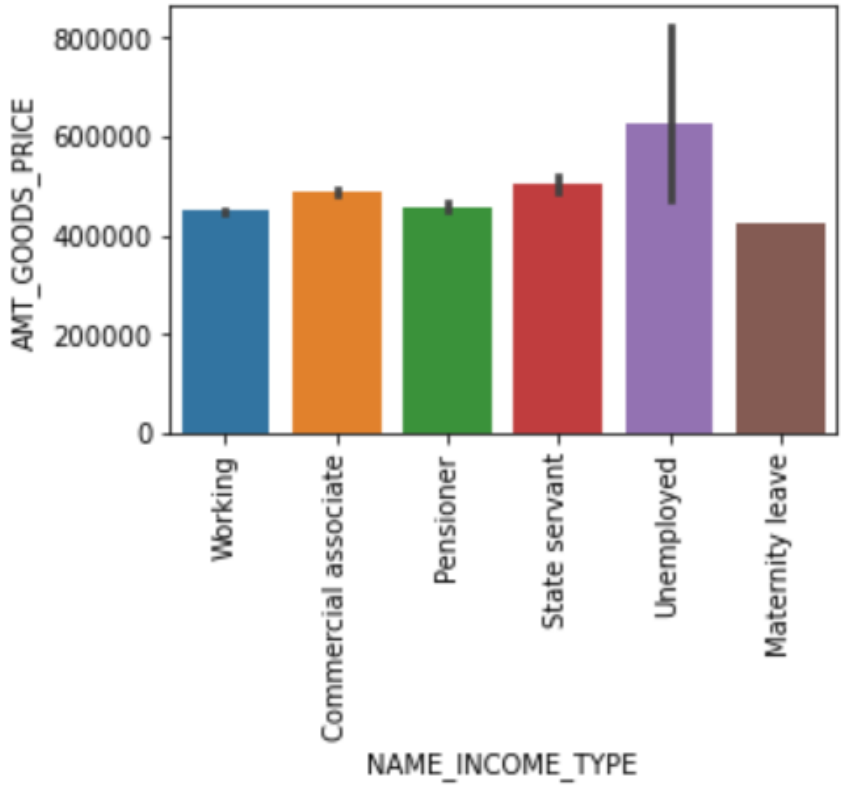


INCOME_TYPE + AMT_GOODS_PRICE

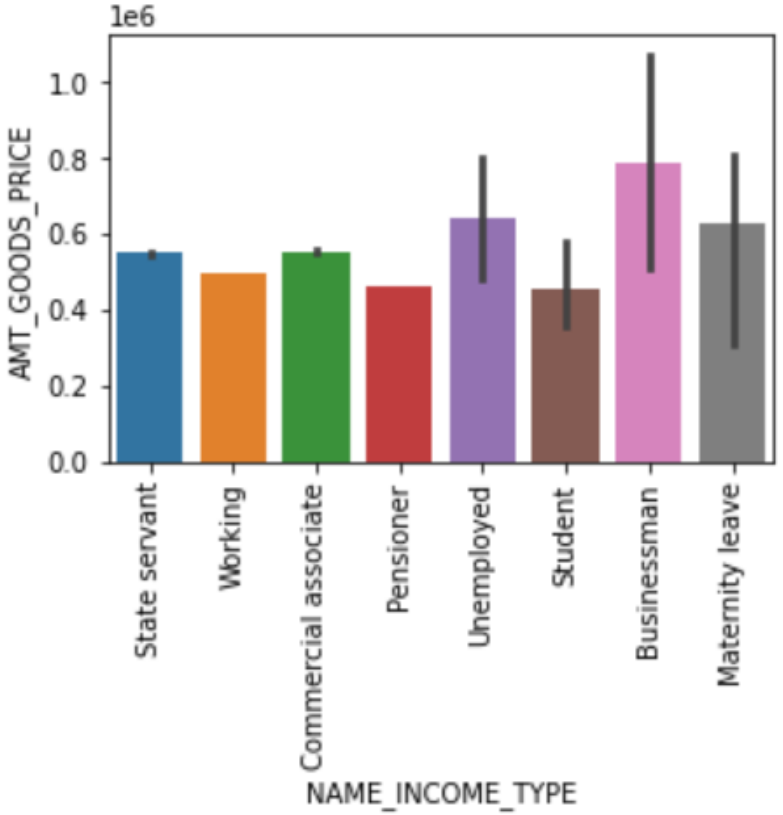
Inference:

Unemployed people in the defaulters cat, even if their count is very low but the goods they applied the loan for seems to be priced the highest. On the other hand businessmen make apply for high priced goods and make payments on time and this category is completely missing in defaulters.

Defaulters:



Non-Defaulters:

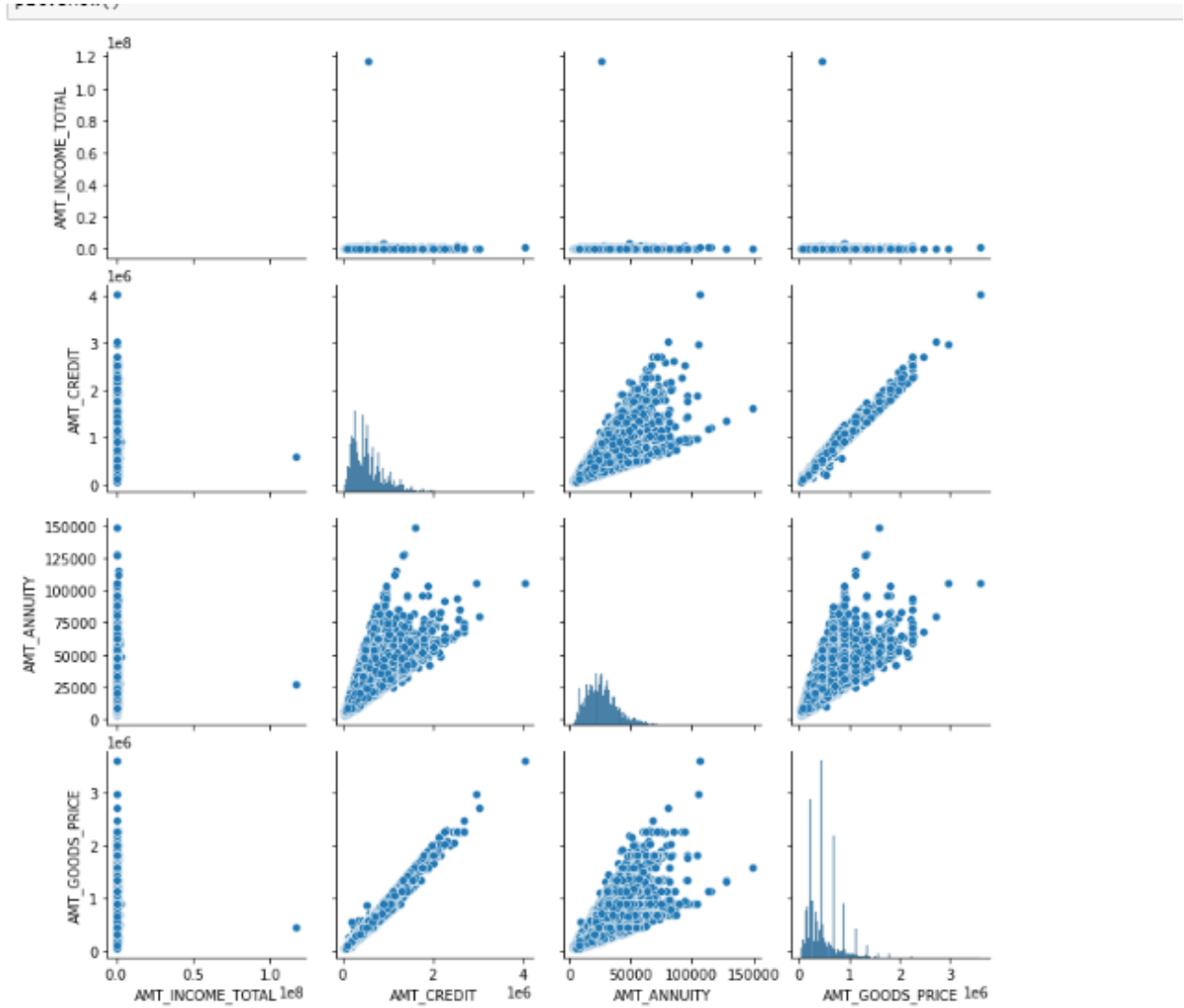


PAIR PLOTS

Inference:

Credit amount and goods price have a strong relationship.

If one increase the other also increases.

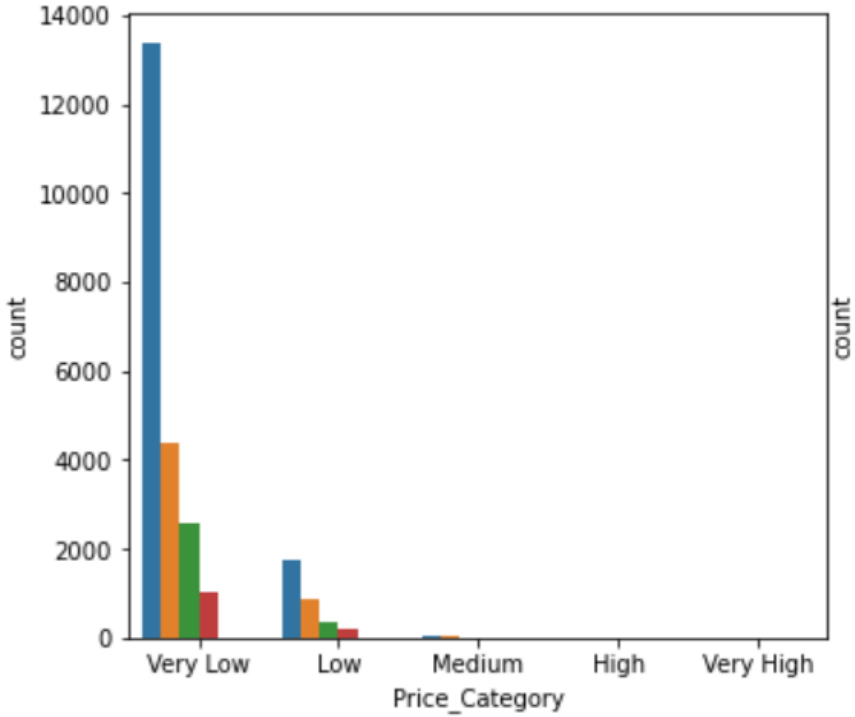


Price_Category-Very Low (Goods Price) + NAME_INCOME_TYPE

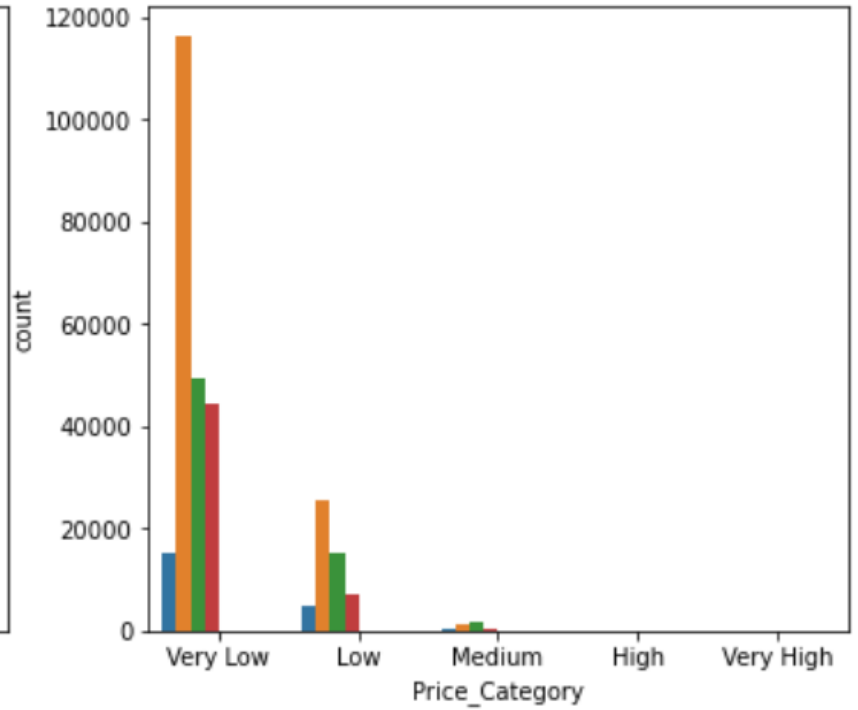
Inference:

Pensioners, Commercial associates and Working pay on time. State servants defaulters.

Defaulters:



Non-Defaulters:

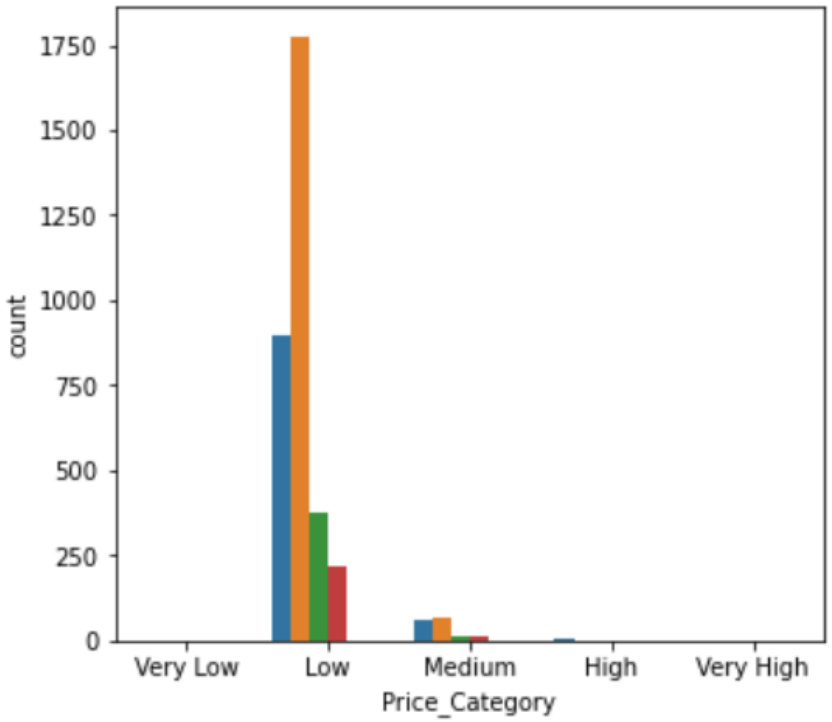


Price_Category-Low (Goods Price) + NAME_INCOME_TYPE

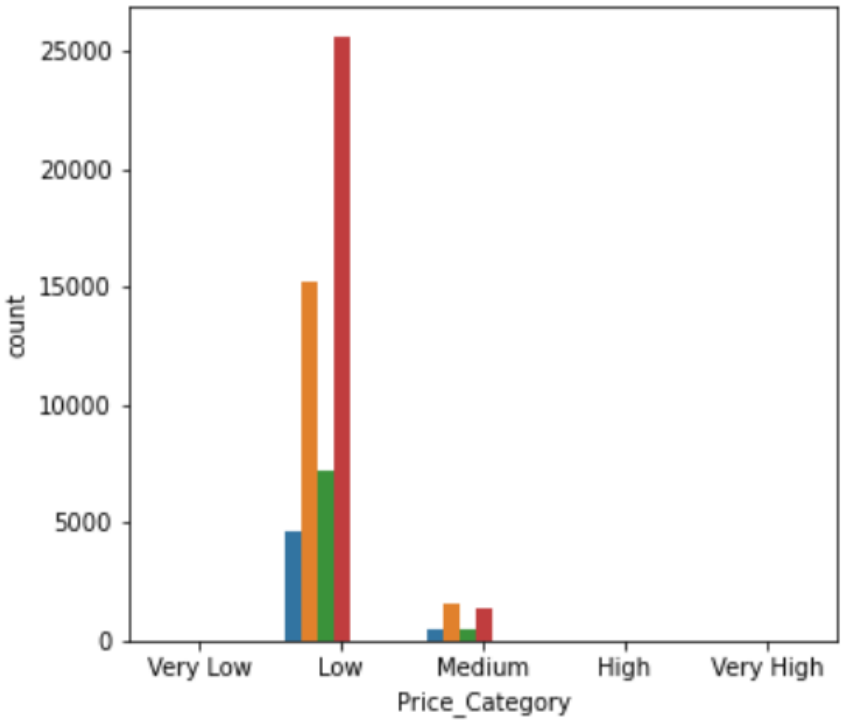
Inference:

Commercial Associates and State Servants - defaulters. Pensioners, Working pay on time .

Defaulters:



Non-Defaulters:

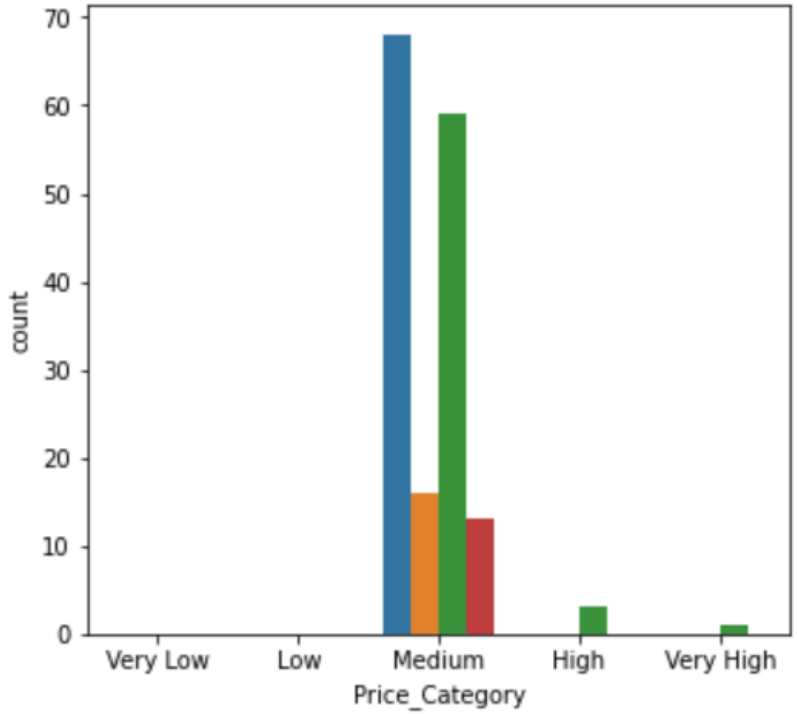


Price_Category-Medium (Goods Price) + NAME_INCOME_TYPE

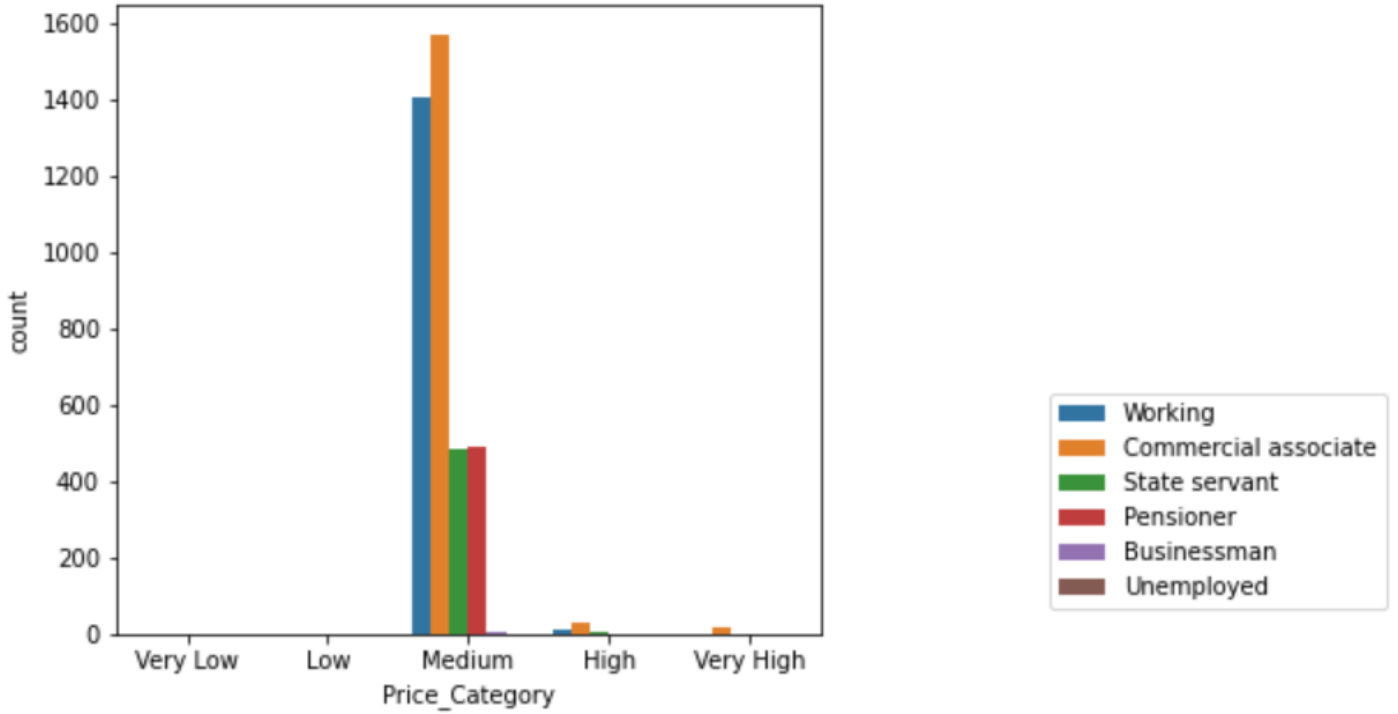
Inference:

Commercial associates and pensioners pay on time. Working and state servant default.
)

Defaulters:



Non-Defaulters:



Final Analysis based on Price Category and Income Type

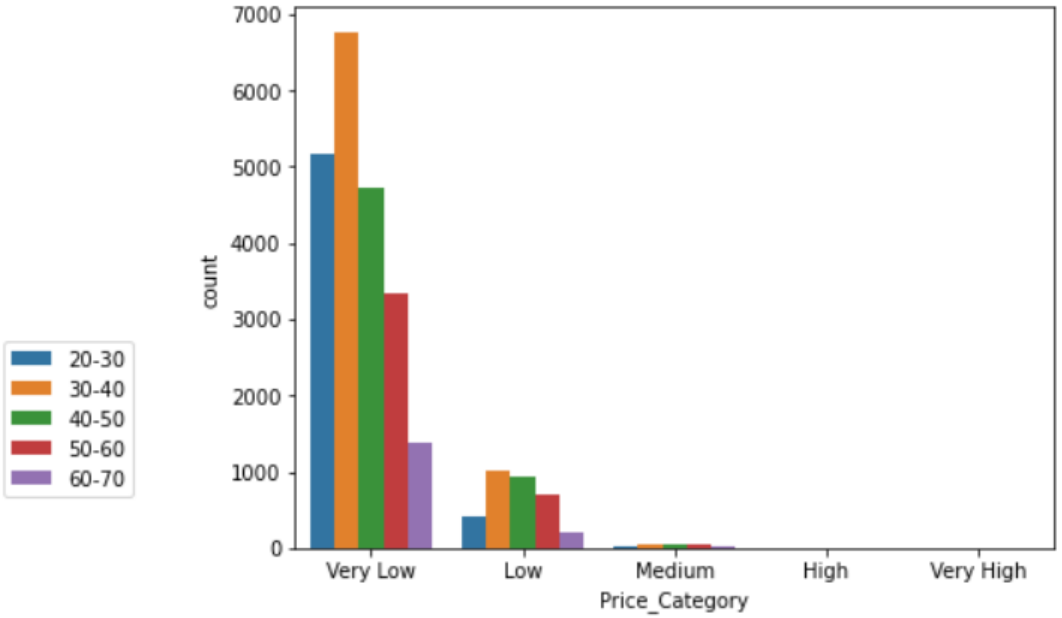
Pensioners and Commercial Associates make on time payments overall.

Age_group + Price_Category

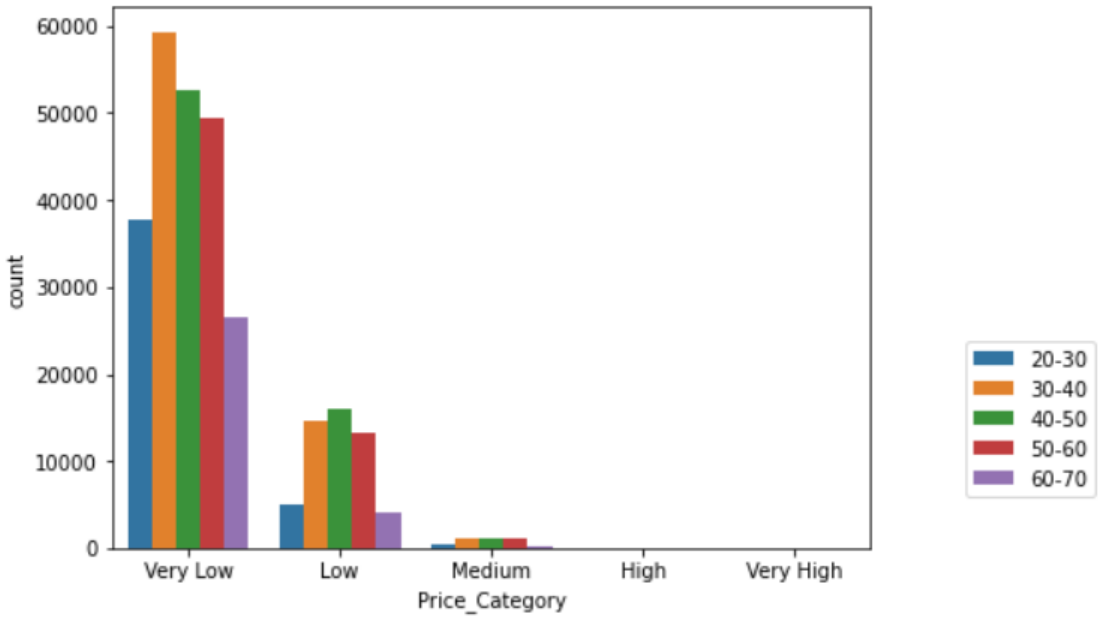
Inference:

Age group 40 and above seems to be makes payments on time across different categories of goods price.

Defaulters:



Non-Defaulters:



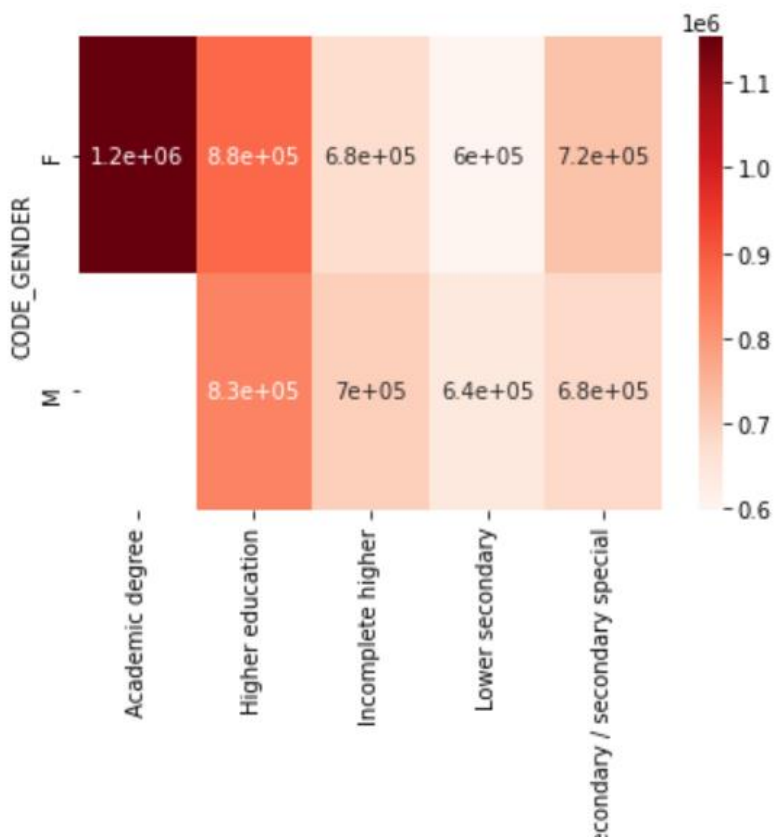
MULTIVARIATE ANALYSIS

ANALYSING GENDER, EDUCATION TYPE AND CREDIT AMOUNT

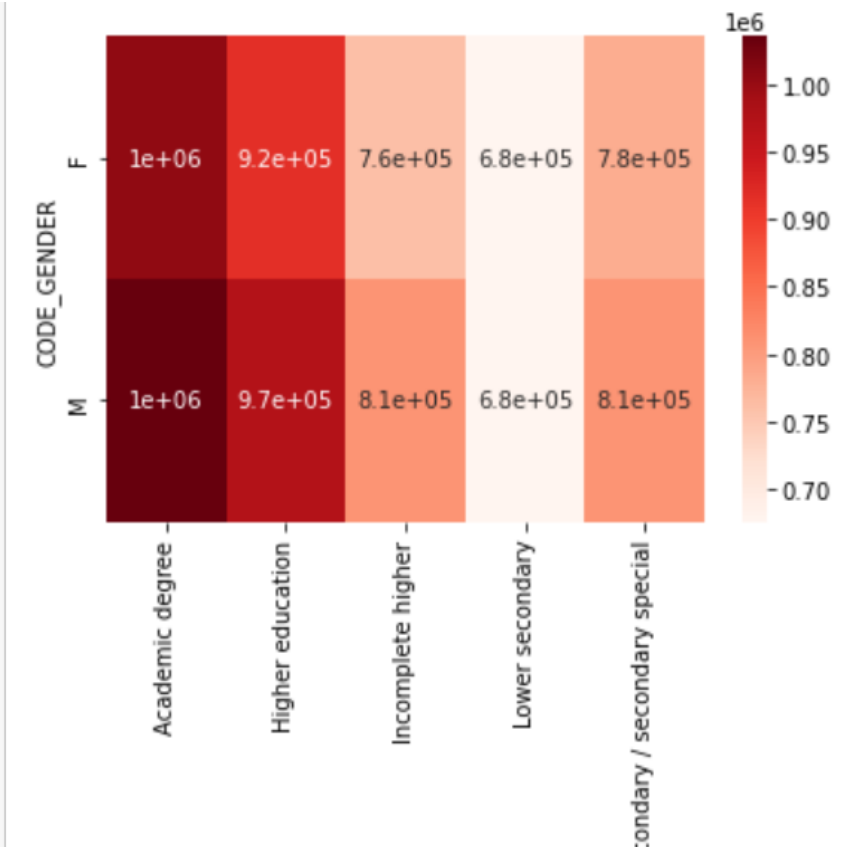
Inference:

Females with academic degrees take high loan credit amounts and defaults the most. Academic Degree holder males and Higher Education Male category seems to be paying on time.

Defaulters:



Non-Defaulters:



Details of second Dataset (previous_application.csv)

1. Shape – **(1670214,37)**
2. Column count with more than 40% null values – **11**
3. Column count with null percentage <40% - **5**

Deleted all columns with null percentage more than 40% as it is a significantly huge number.

Analysing columns with null percentage >0<40

1. AMT_GOOD_PRICE – 23%

There's a noticeable difference between mean and median which implies there are outliers. Also mode value is 45000.0 but since the count of null value is very high it is better to leave them as it is and not impute as it may hamper the analysis.

2. AMT_ANNUITY – 22.2%

There doesn’t seem to be any valid reason for these missing values. The mean and median are not very distant in terms of their values but with the null percentage being too high it is better to leave them than imputing them.

3.CNT_PAYMENT– 22.2%

The median and mode values are same and there's not much difference between mean and median so we can impute them with median but since the percentage of null values is high, we can leave these values as it is.

4. PRODUCT_COMBINATION – 0.02%

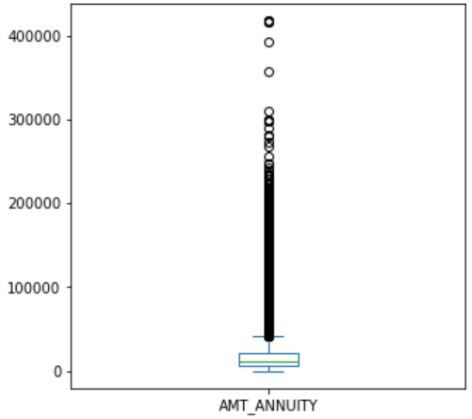
The most recurring category is Cash and since the count if null values is also less it is ok to impute those with 'Cash‘

OUTLIER (UNIVARIATE ANALYSIS)

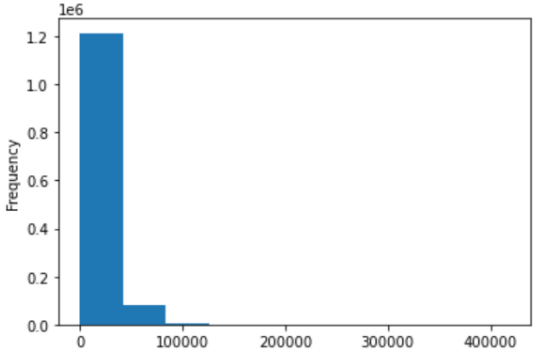
AMT_ANNUITY COLUMN

Inference:

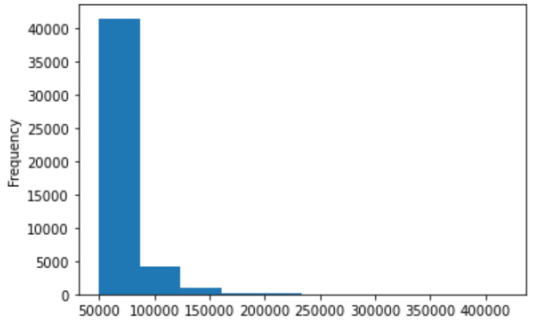
It looks like that even beyond 50000, which is around the 95th percentile there is a huge chunk of high values which is tightly attached to the upper fence. This chunk could be of people who are well off. So all the values beyond the 99th percentile are outliers and they should be ignored while doing analysis.



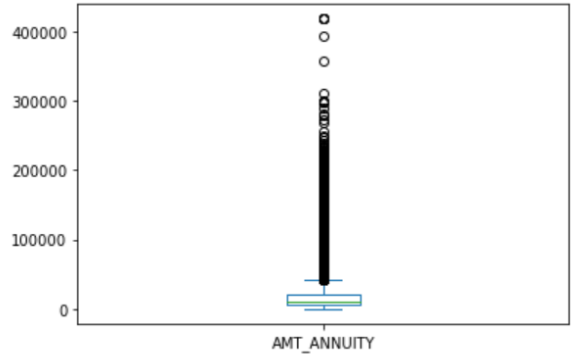
First plot



First histogram



Histogram for values above 50000

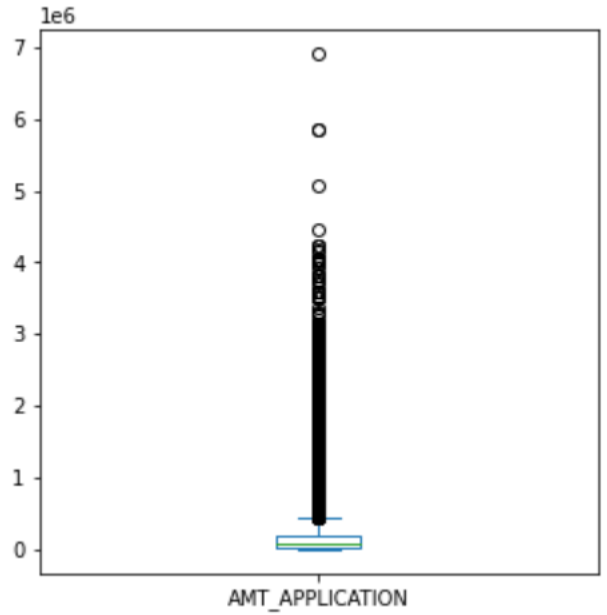


Plot for values<800000

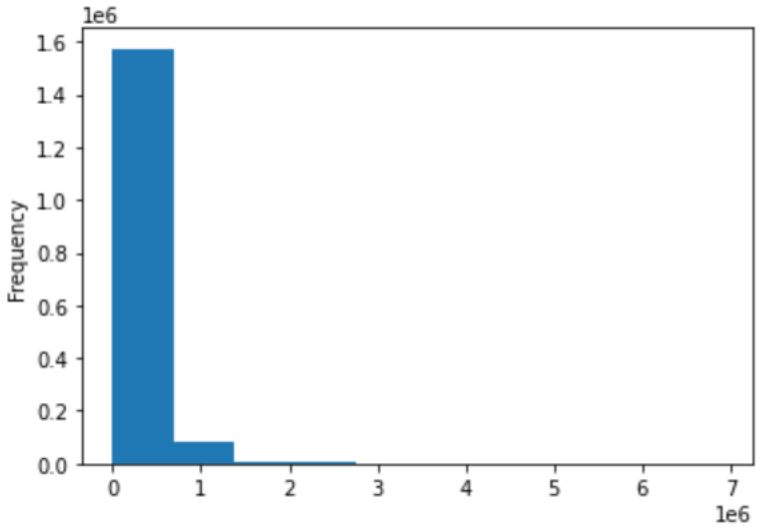
OUTLIER (UNIVARIATE ANALYSIS)

AMT_APPLICATION COLUMN

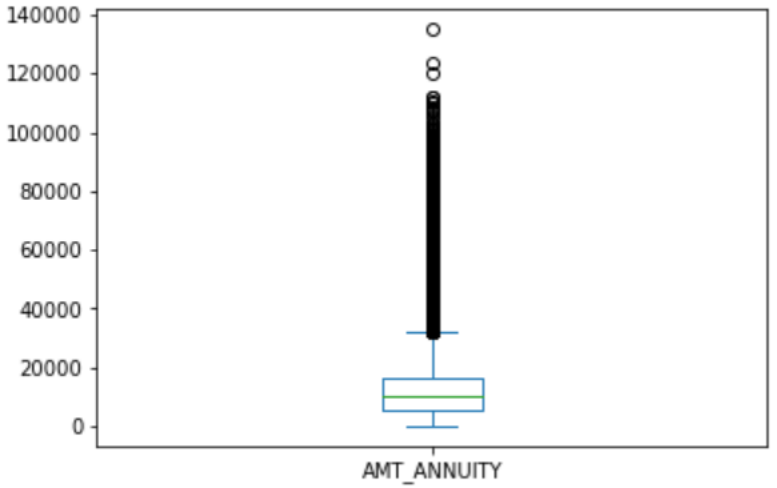
Outliers beyond 90th percentile. Should not be considered during analysis.



First plot



First histogram



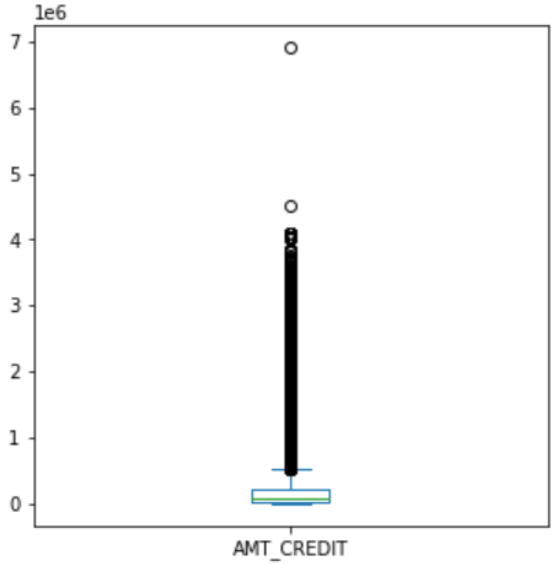
Plot for values < 500000

OUTLIER (UNIVARIATE ANALYSIS)

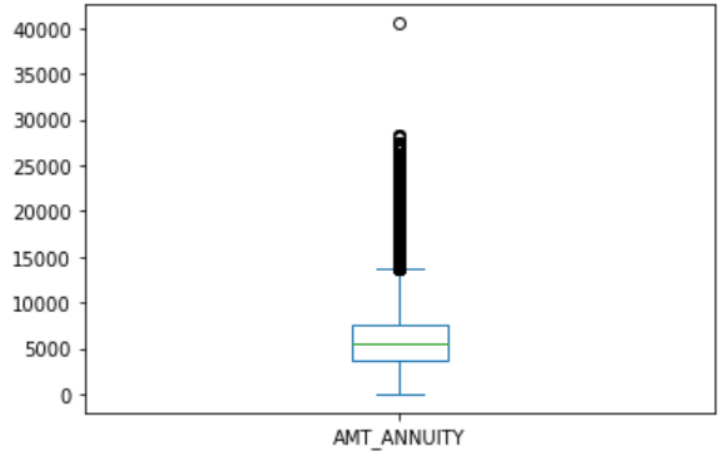
AMT_CREDIT COLUMN

Inference:

There's huge jump between the 50th and the 75th percentile. Values beyond 504805.5 are outliers in this case.



First plot



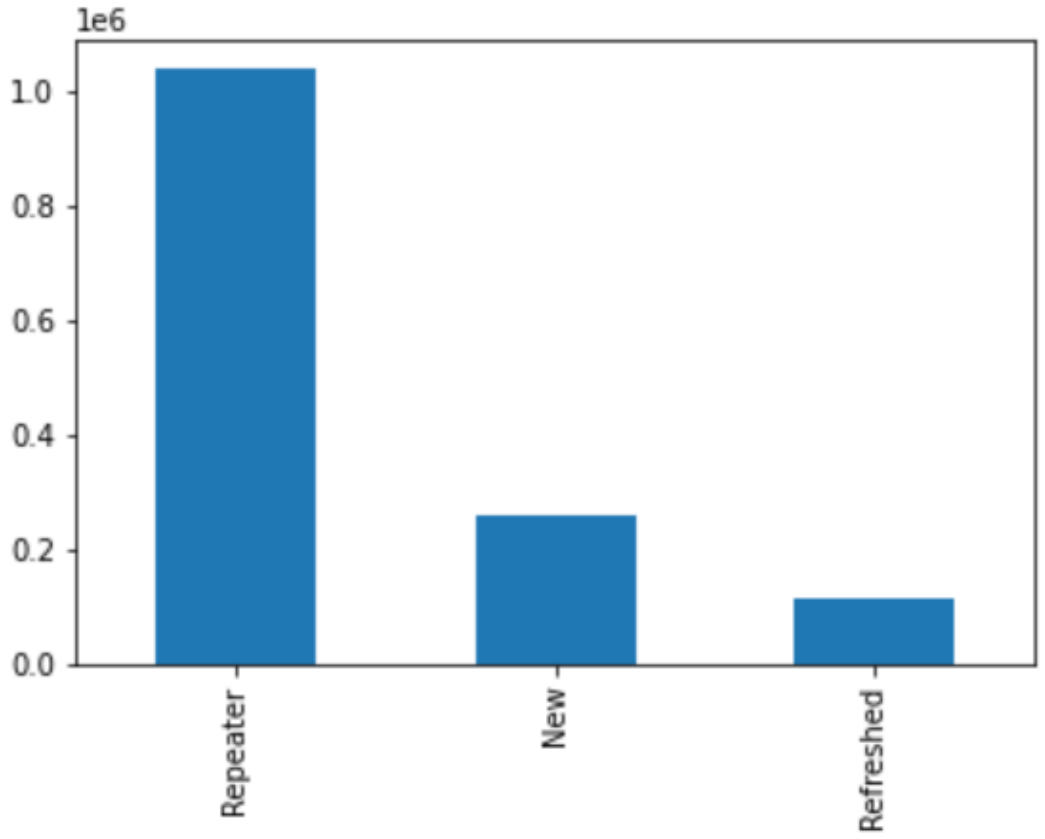
plot for values < 80000

ANALYSIS AFTER MERGING TWO DATAFRAMES

CLIENT TYPE

Inference:

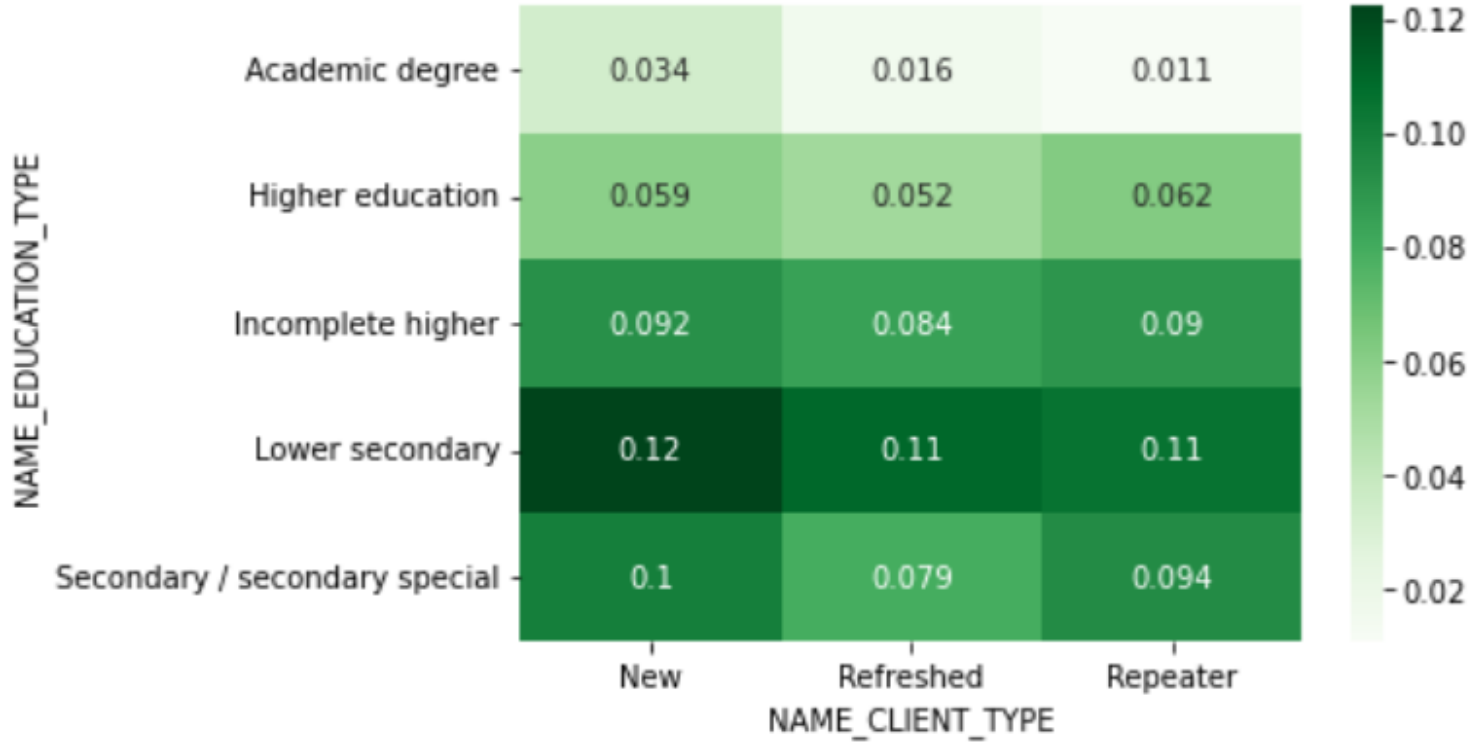
Most of the applicants are repeat.



NAME_EDUCATION_TYPE and NAME_CLIENT_TYPE based on mean of TARGET variable.

Inference:

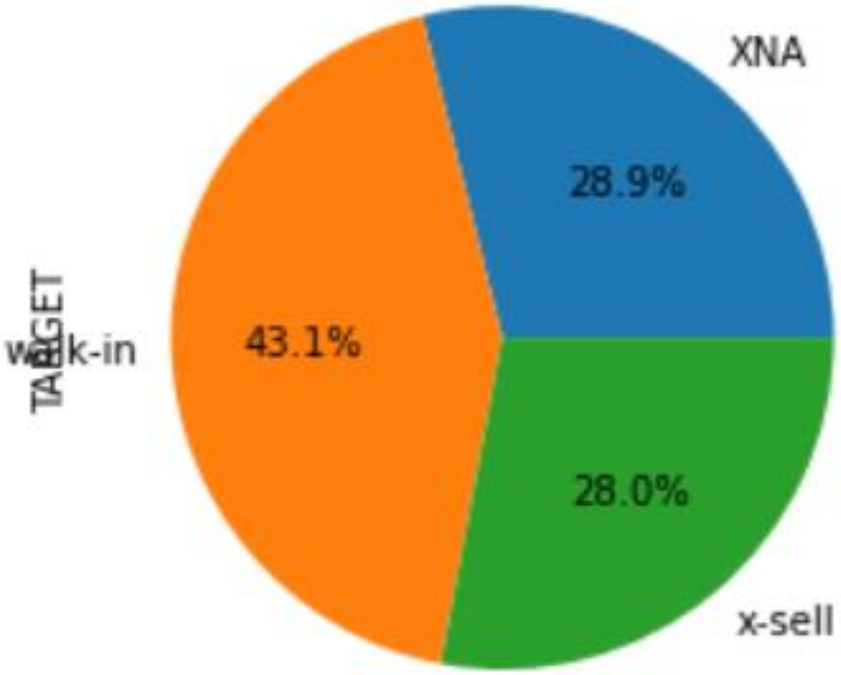
Lower secondary education cat and New clients seem to be defaulting the most. Repeat clients with Academic degrees pay on time followed by Refreshed clients with academic degree. Also Academic degree followed by Higher Education people seem to be making payments on time.



NAME_PRODUCT_TYPE based on TARGET variable

Inference:

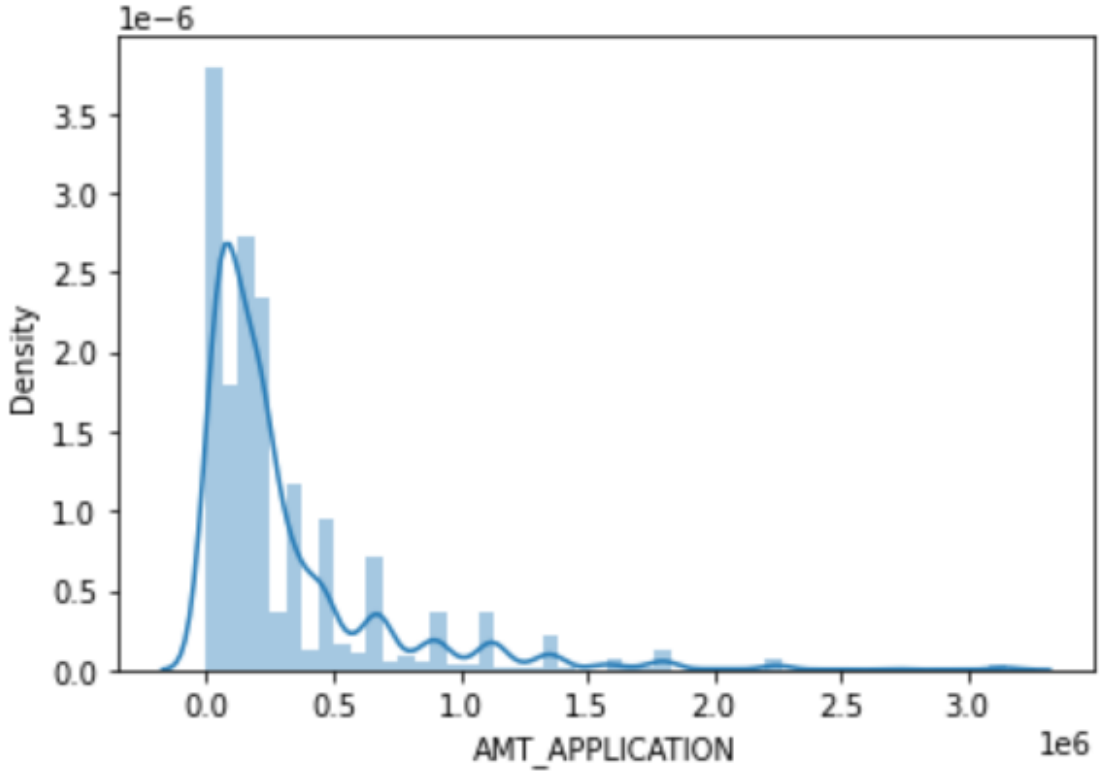
XNA and x-sell category applicants pay on time compared to walk-in applicants



AMOUNTS APPLIED BY WALK-IN CLIENTS

Inference:

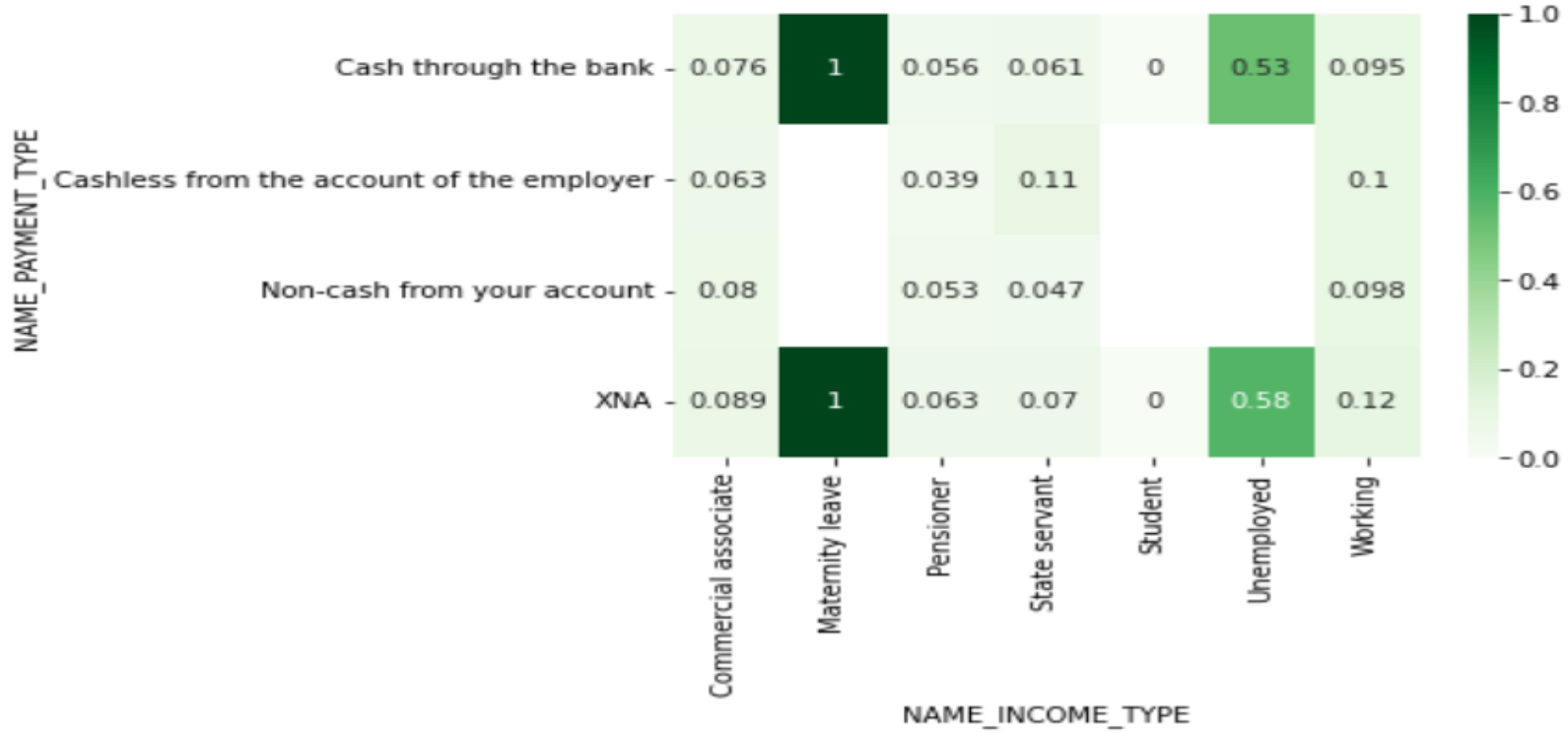
Walk-In clients applying for small amounts for loan default the most.



NAME_PAYMENT_TYPE and NAME_PAYMENT_TYPE based on mean of TARGET variable

Inference:

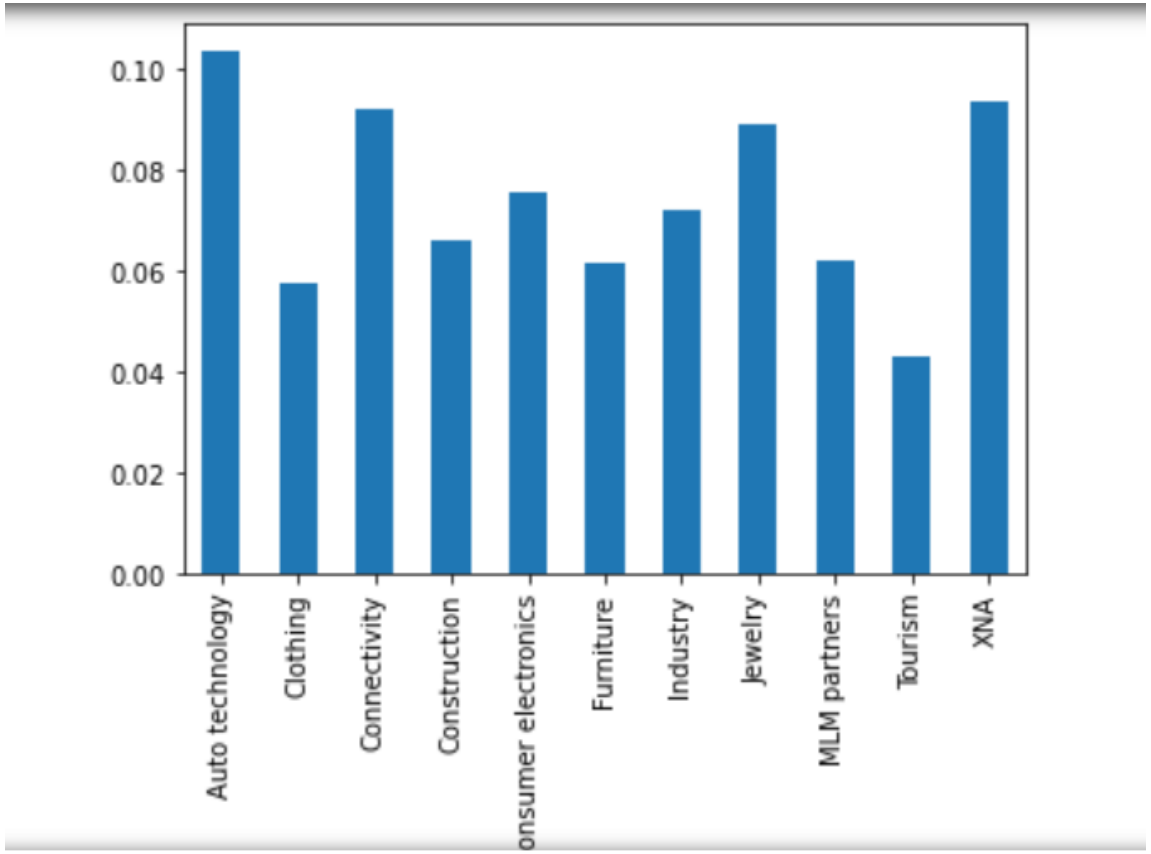
Pensioners who choose payment type as Cashless from the account of the employer pay on time.



Industry Type based on TARGET variable

Inference:

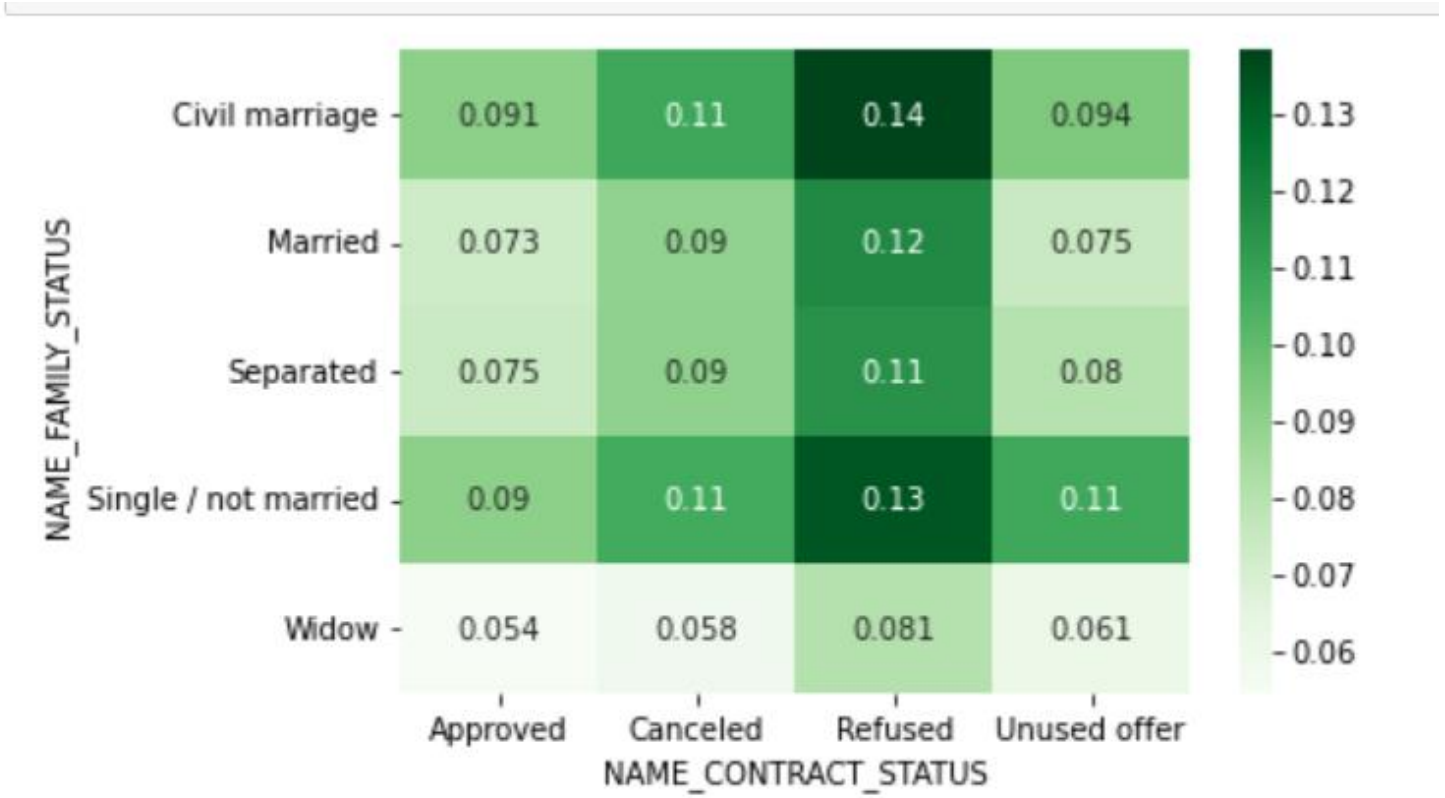
Auto Technology sector has most defaulters. Tourism has the least.



NAME_FAMILY_STATUS and NAME_CONTRACT_STATUS based on mean of TARGET variable

Inference:

Widows overall and Married applicants whose previous loan was approved

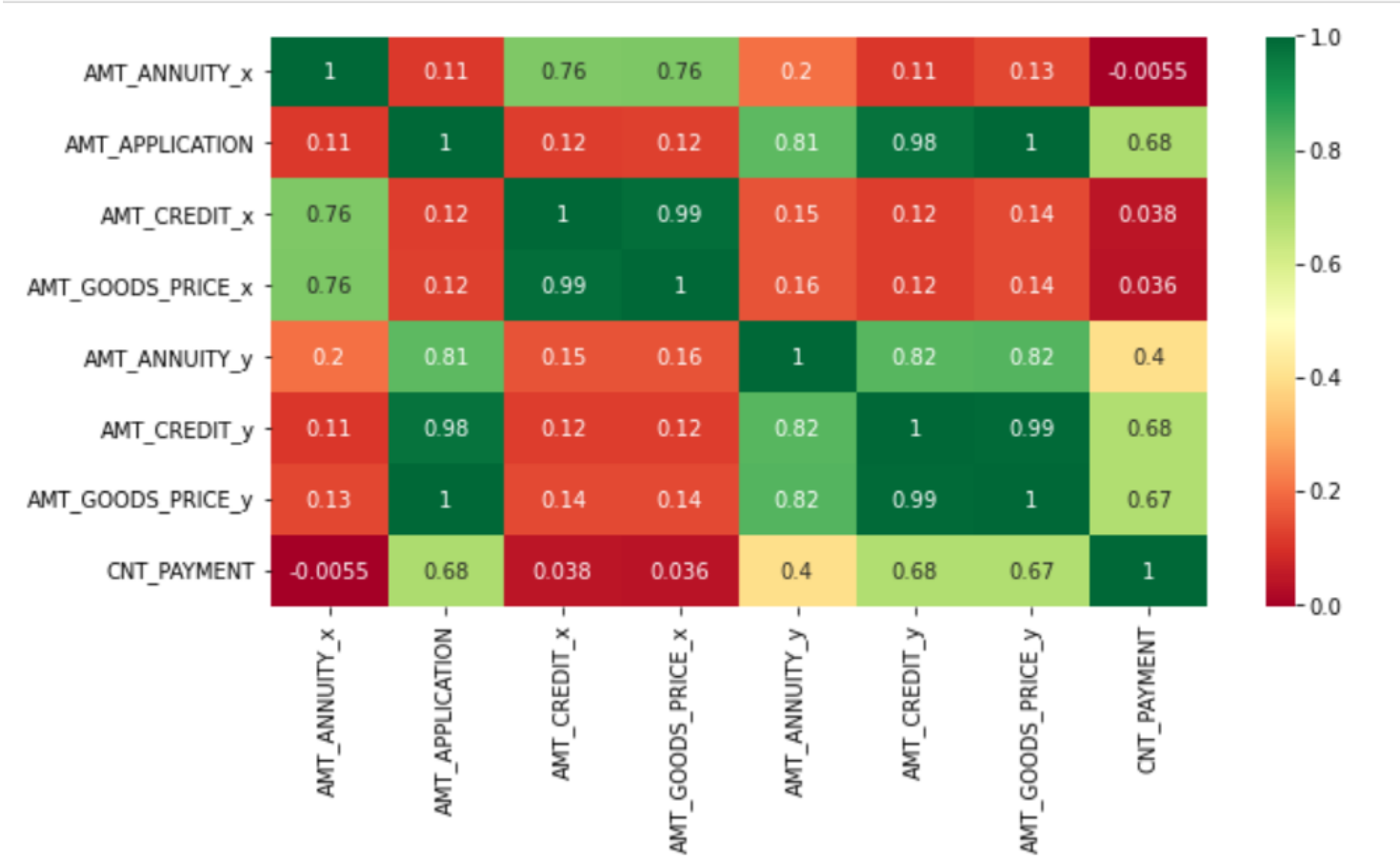


CORRELATION MATRIX

Inference:

The strongest correlations:

1. **amt_application** and **amt annuity_y** = **0.81**
2. **amt_application** and **amt_credit_y** = **0.98**
3. **amt_credit_x** and **amt_goods_price_x** = **0.99**
4. **amt_annuity_y** and **amt_credit_y** = **0.82**
5. **amt_annuity_y** and **amt_goods_price_y** = **0.82**
6. **amt_credit_y** and **amt_goods_price_y** =**0.99**



FINAL CONCLUSION

AUDIENCE TO TARGET:

1. Clients with higher work experience, 15 years and above have higher incomes and make payments on time.
2. Clients with 0-1 children.
3. Commercial Associates and Pensioners.
4. Males with Academic degrees followed by Higher Education.
5. Married and Widows.
6. Age-Group 40 and above.
7. Clients with score 0.5 and above.
8. Businessmen.
9. Seller Industry – Tourism
10. Repeater clients.